THE UNIVERSITY
of EDINBURGH

Theory of mind and language evolution: an exploration of rapid and involuntary perspective-taking

Cathleen Jeannette O'Grady

Doctor of Philosophy
The University of Edinburgh
2020

Declaration

I declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Material presented in Chapter 2 was previously published as a chapter in the *Oxford Handbook of Psycholinguistics* by Cathleen O'Grady (author of this thesis and declaration) and Kenny Smith (PhD supervisor). This chapter was conceived and edited by both authors. I carried out the research and writing.

An early description of the research and initial analyses presented in Chapter 4 appeared in the Proceedings of the 39th Annual Meeting of the Cognitive Science Society. This proceedings paper was co-authored with Thom Scott-Phillips, Suilin Lavelle, and Kenny Smith (PhD supervisors), and has been reproduced here with the permission of all authors. The research was conceived by all authors, and all authors contributed to the writing and editing of the paper. I conducted the research, and analysed the data with assistance from Kenny Smith.

Material presented in Chapters 4 and 5 has been submitted to the *Quarterly Journal of Experimental Psychology*. This paper was co-authored with Thom Scott-Phillips, Suilin Lavelle, and Kenny Smith (PhD supervisors), and has been reproduced here with the permission of all authors. The research was conceived by all authors, and all authors contributed to the writing and editing of the paper. Data for Experiment 5 in Chapter 5 was collected by Rachel Kindellan. I conducted the research, and analysed the data with assistance from Kenny Smith.

Cathleen Jeannette O'Grady

Lay Summary

Human language is an incredible phenomenon. I can take an infinite array of ideas from my own mind and transfer them to someone else's mind, just by combining sounds or handshapes into words and words into sentences. Other animals have fascinating and sophisticated communication systems, but the ability of human language to convey infinite meanings is unique.

How do we explain the features of language, and why do only humans have it? Researchers have approached this question using a range of different techniques. They've built computer programmes that test out how "bots" can learn language from each other, and what happens to those languages over time. They've tracked how new languages, like Nicaraguan Sign Language, emerge. And they've taught human participants toy languages, observing how the process of learning and using a language changes the features of that language.

This research has taught us a lot about how language is shaped by cultural learning, but it doesn't capture something important about how language works in the real world. In a language experiment in a lab, participants are given a word in a toy language (like "blorp") and taught what it means (like "a blue circle"). So, they're given both the *signal* and the *meaning*. Bots in computer simulations are also given both the signal and the meaning. But newly emerging languages, and babies learning language, have to make this link between signal and meaning for themselves.

Humans seem to be really good at this: I can have a pretty good guess at what a brand-new signal means, even if it's a really strange one. For instance, if my friend pulls a weird face at me from across the room, I can draw on the context to try to figure out what she means. This ability to figure out the meaning of a new signal, and to know when someone's behaviour suggests that they're trying to communicate (rather than, say, hold in a sneeze), is a really important component of understanding how communication systems emerge.

To make these links, people need to be able to think about what other people know, think and believe. If I want to understand what my friend is trying to tell me when she pulls a face, I need to think about what's going on in her head. I need to do this even when she uses perfectly normal language, too – if she asks me "How did it go?" I'll need to draw on our mutual knowledge to

understand what "it" means. This suggests that language relies on *mindreading*, which in this case doesn't refer to any psychic abilities, but rather the everyday ability people have to think about the contents of other people's minds. The problem is, it's not very clear how good people really are at mindreading. Some evidence suggests that children only develop it after they've already learned language, and that it's pretty tough for adults. Other evidence suggests that we do basic mindreading (just keeping track of what other people can see) so automatically that we don't even realise we're doing it.

In this thesis, I take a close look at some of the evidence for automatic mindreading. There's an experiment that initially seemed to show that people keep track of what other people can see without thinking about it at all, but later research using a similar experiment suggests that maybe it's not that simple. A big problem with this research is that different researchers use slightly different versions of the experiment, and so they get results that contradict one another. I ran a series of experiments making small modifications to the basic task to try to understand why there are so many contradictory results, and to get a better handle on automatic mindreading.

I find that people do seem to keep track of what other people can see, and we do it really fast and without thinking about it at all. But we only do it when there's some reason to think about those other people's perspectives – that is, we don't track someone's perspective just because they're there, but rather only if we have reason to think their perspective is important. So, this kind of basic mindreading is more like seeing in *focus*, which requires some part of your brain to think about what's important to look at, and less like seeing in *colour*, which just happens all the time. But my research also suggests that to research these questions properly, we need much bigger experiments that very carefully keep track of tiny changes to the task.

Finally, I come back to the question of whether people really need mindreading to make links between meanings and signals. I suggest that there's a different psychological phenomenon that might be more helpful to understanding this corner of language evolution: the idea of joint attention, which is what happens when two people are both focused on the same thing, and both know that they're both focused on the same thing. When I'm playing a game with a friend, we're in a state of joint attention to the game. I suggest that joint attention could be a better way to think about people's communication skills than mindreading, and point out some possibilities for future research.

Abstract

Current research on language evolution has provided considerable insight into the emergence of linguistic structure, but much of this research does not account for how language users access speaker meanings to build mappings between meanings and signals. Signal-meaning mappings may arise in natural communication systems by three main routes: signal first, meaning first, or signal and meaning simultaneously. Humans seem uniquely capable of creating new simultaneous signal-meaning pairs as a result of our ability to make our communicative intentions apparent. This "ostensive-inferential" communication can explain how meanings are attached to novel signals in infancy, emerging languages, and everyday communication by appealing to mindreading, or theory of mind – the ability to infer the intentions, knowledge, and other mental states of interlocutors.

The ostensive-inferential model of communication has been criticised on the grounds of implausibility; specifically, that neither infants nor adults are capable of the kind of rapid and complex mindreading required by this model. However, a range of experimental paradigms appear to show evidence of rapid and unconscious mindreading in both adults and infants. One paradigm in particular, the Dot Perspective Task (DPT), has been argued to show evidence of "spontaneous" or "automatic" mindreading in adults, although this interpretation is the subject of considerable dispute, with some evidence suggesting that the results are best explained by cognitive processes that do not involve mindreading.

In this thesis, I present a range of experiments using the Dot Perspective Task to investigate whether the classic result in this paradigm is best explained by mindreading or by alternative explanations. Using an adapted and extended set of DPT stimuli, I first investigate whether the "mindreading" effect is found only for human-like avatars, or whether it extends to non-human arrows. An exploratory analysis of the data from this task suggests that the results may be best explained by a spatial confound in the stimuli.

A subtle but critical difference in the implementation of DPT variants in the literature may explain not just this result, but other irreconcilable differences in the DPT literature. In a se-

quence of five experiments, I investigate this variation in experimental method. The results from these tasks suggest that the DPT does not demonstrate automatic mindreading – that is, mindreading that is reflexive and purely stimulus driven; but rather spontaneous mindreading, or mindreading that is rapid, unconscious and involuntary, but directed by attentional systems. This finding prompts us to clearly distinguish spontaneity from automaticity.

I present two further experiments investigating the mechanism underlying the classic DPT result. An adaptation of the DPT produced null results, as did a simplified version of the same task. This null result may be explained by the increased demand of the task, but given inconsistent findings across the DPT literature, it may also be the case that the classic DPT finding is less robust than it appears to be. I therefore review the current literature on failed replications in psychology and other disciplines, identifying how the problems described in this literature are relevant to the Dot Perspective Task. I argue that the use of the DPT to investigate rapid and involuntary mindreading remains promising, but that drawing firm conclusions from this paradigm will require replications of the current results with sufficient statistical power, as well as a thorough investigation of the wide range of methodological implementations that may affect the results of the task.

Finally, I explore an alternative route for future investigation of the plausibility of the ostensive-inferential model of communication. I discuss current attempts to reconcile the ostensive account with empirical research on mindreading, arguing that "minimalist" accounts of ostension fail to capture some of the central features of ostensive communication. I then suggest a reformulation of ostension that draws on the concepts of joint attention, common ground and mutual manifestness to offer an account of ostension that is both developmentally and cognitively more plausible than an account that relies on mindreading. I suggest that joint attention would be a fruitful avenue for future research on ostension.

Acknowledgements

My PhD was funded by the University of Edinburgh's Principal Career Development Scholarship and Global Research Scholarship. I am grateful not only for this financial support, but also to my PhD supervisors for their support in securing funding. Thanks also to all the people who helped me make the leap from Rhodes to Edinburgh: the Skye Foundation for funding my MSc, and all of my very many wonderful lecturers at Rhodes (particularly Francis, Ian, Ralph and Tom) for their encouragement, mentorship, and countless reference letters.

I could not have asked for three better supervisors as I careened through four years during which life outside the ivory tower came at me thick and fast. Thanks to Kenny for his humour and boundless expertise, and for never rolling his eyes when he unpicked in two minutes a coding issue that had been giving me headaches for days; to Suilin, for constructively guiding me through treacherous theoretical waters and infecting me with enthusiasm for beaver rewilding; and to Thom, whose encouragement got me to the starting line of the PhD as well as through it, whose work has left me constantly marvelling at ostension, and who has a gift for making me somehow believe that I might have something reasonable to contribute.

The wider CLE has been a wonderful environment in which to spend these years. Thanks to all of you for the beach days, coffee mornings, and thoughtful insight, and particularly to Simon, Jenny, Marieke S. and Chris for their valuable critiques and discussions of my research. To my fellow CLE PhD students – Andres, Ash, Carmen, Fang, Fausto, Fiona, Jasmeen, Jon, Jonas, Kevin, Marieke W., Svenja, Tamar and Yasamin – thanks for the camaraderie and skill-sharing, and for the feasting and hilarity at the PhD retreats. Particular thanks to Carmen and Marieke for rescuing me from all the plumbing disasters, and to Marieke (again) for being my thesis guide. Thanks also to Rachel for swooping in to collect data at the eleventh hour.

I'll greatly miss the wonderful office environment of 1.15; thanks to everyone who made it such a great place to work, and especially to my "cohort+" gang: Candice, Ernisa, Eva, Jade, Louis, Maki, and Wenjia. Thanks especially to Jade for the political rants and gear advice, Maki for cosy chats and being the social glue, and Eva for being such a fun and kind office buddy and now out-

of-office buddy too. Thanks also to the unflappable and incredibly kind Postgraduate Office staff and IT staff, who make the Dugald Stewart Building such a delightfully friendly environment, and without whom everything would crumble.

My science writing work has introduced me to more wonderful people than I could ever name, but special thanks must go to my editor John, for opening countless doors for me and for his unerring patience while I had far too many balls in the air. Thanks again to my supervisors too, for bearing with me during the juggling act. I am probably their first PhD student to hit the pause button to travel to South Africa for a rhino dehorning; I hope for their sake I will be the last.

My friends outside the academic bubble – you have saved my sanity. To the scatterlings and all those at home in SA, thank you for the Skype dates and visits. To the Phoenix Choir, even though my tenure was brief, thank you for the joy and levity, and especially to Alex F., Carla and Giada, who bounced into my life very suddenly with bursts of treasured friendship. Thanks to Ivet for always making me feel like I have family living right around the corner; and to Alex B., for the food, tea, sounding board, and for generally being one of the best people I have ever known. And Kati, my *other* other half, and the one stable table leg – there is really too much to say. Thank you for everything you are, and everything you have been all these years.

I am so grateful to my family for their support and encouragement. Dragane i Ljubo, hvala puno za preukusnu hranu, poklončiće, ljubav, i vašeg divnog sina. Stephen, Karlien, and my darling Eirin and precious Morgan – thanks for the distraction, cuteness, raucous games, and beautiful drawings. Dad, thank you so much for all the walks, pit crewing, roadtrips, long chats, blue eggs, and for being such a wonderful friend and advisor. Thanks also for hosting the writing retreat – I quite literally could not have done this without you feeding and caffeinating me through the fugue state. And my wonderful Nik, I would need at least another 100,000 words to properly account for all your halp [sic], but: thank you for fixing all my crises, for standing back-to-back with me against the orcs, and for the countless joys of these twelve years. Obožavam te, moj šnekiću.

Finally, thank you to my incredible mother, whose mischief, humour and love still echo through everyone she left behind; whose unerring delight and curiosity in the world around her undoubtedly set me on the path of extreme nerdery; and who helped me through every single day of this PhD, even after she was no longer with us.

Contents

Chapter 1

Thesis outline

Everyday social interaction requires inferring the intentions, knowledge, and other mental states of the people around us. In order to invite a friend to join a group dinner, I must first be aware that she does not yet know the dinner is being planned, and also have an awareness of her social network and preferences – for instance, whether she knows and likes the other people coming to the dinner. If a pedestrian walking in the bike lane is oblivious to me approaching on my bike from behind, I ring my bell to alert them to my presence, which requires me to have an understanding of what they do and don't know about their immediate environment.

This kind of understanding – referred to variously as mindreading, mentalising, and theory of mind – plays a crucial role in language. Utterances like "I'm tired" could, in different contexts, mean "Let's go to bed" or "Let's turn back from this walk" – ascertaining the meaning requires not just understanding the words, but also the speaker's intentions given the context. Despite the critical role of this pragmatic competence in language, it has received minimal attention from research on language evolution (Scott-Phillips 2015).

This thesis explores the role of mindreading in language evolution, focusing on an experimental paradigm that investigates rapid and efficient mindreading in adults. In Chapter 2.1, I survey the current literature on evolutionary linguistics, focusing on the insights that have been gained from studying language as the product of a dual inheritance: the genetic inheritance that enables the cognitive capacities underlying language, and the cultural inheritance that passes a particular language from one generation to the next. This work has focused on the emergence of linguistic structure, leading to a strong body of evidence showing that the various pressures operating during cultural transmission can explain the emergence of this structure.

In Chapter 2.2, I discuss the gap that this literature leaves around mindreading in language evolution: how language users infer the meanings of signals. Much of the research on language

evolution provides pre-established signal-meaning mappings to human participants (or computational agents). However, in the naturalistic settings relevant to the emergence of linguistic systems, people must independently map the referents and intentions of speakers to objects and events in the environment. People are routinely faced with situations in which they must interpret novel signals that are delivered without an established communication system: a driver flashing his headlights in an attempt to deliver a warning; a friend raising an eyebrow meaningfully without saying anything; understanding the goal of someone pantomiming a request for water in a situation with no shared language; acquiring the sign language used by one's peers at school; acquiring language during infancy. The ability to decipher the meanings of these signals is dependent on being able to infer the intentions of a speaker. I describe an account of human communication that gives central importance to this expression and interpretation of intentions: the *ostensive-inferential* model. This model accounts for the human ability to infer the communicative intentions of others by appealing to advanced, flexible and efficient mindreading.

The evidence on whether people are in fact capable of such advanced, flexible, and efficient mindreading is mixed. There is substantial debate on what constitutes true mindreading, who reads minds, and when they do so. Because of this, the plausibility of the ostensive account has been questioned (see e.g. Moore 2014). Chapter 3 surveys the current evidence on mindreading in typically and atypically developing children and in adults; across human cultures; and across species. I discuss points of debate in theoretical approaches to mindreading, focusing on nativism vs constructivism, one-system vs two-systems accounts, and mentalising vs submentalising accounts. I argue that the ostensive account depends on a one-system, mentalising account of mindreading, and that empirical work testing the predictions of these accounts may therefore be informative in assessing the plausibility of the ostensive-inferential model. I survey research that explores these questions by focusing on rapid and involuntary mindreading in adults, and describe the Dot Perspective Task (DPT), the paradigm that is used in the experimental work in this thesis.

Chapters 4, 5 and 6 present a series of experiments using the DPT to test predictions made by the mentalising/submentalising and one-system/two-systems accounts of mindreading. Chapter 4 focuses on the mentalising/submentalising distinction by testing whether people respond to the "perspective" of arrows as well as to humanoid avatars. The results from this experiment highlight a crucial methodological inconsistency in the DPT literature that appears to explain apparently contradictory results between different implementations of the task.

Chapter 5 presents a series of five experiments that investigate this methodological inconsis-

tency, reconciling the apparent contradictions in the literature and providing evidence for rapid and involuntary mindreading that is nonetheless not automatic – that is, not purely stimulus-driven. Chapter 6 extends this experiment design to investigate the mechanisms underlying this mindreading effect, finding a series of null results. In light of these null results, as well as other null results reported in Chapter 5, Chapter 7 discusses the potential for widespread statistical problems in the DPT literature. In this chapter, I survey the current literature on the replication crisis in the behavioural sciences, identifying how various research methods that contribute to a low replication rate pertain to the DPT literature. I suggest that a research programme strategically manipulating a range of elements of experiment design in the DPT, and using substantially higher-powered samples, is necessary to gain a clearer picture of the triggering conditions and limitations of rapid and involuntary mindreading.

Chapter 8 returns to the ostensive-inferential model of communication. I discuss the implications of the research presented in Chapter 4 to 6, arguing that these results offer tentative support for the ostensive-inferential model, in that they collectively provide evidence against two-systems and submentalising accounts of mindreading, and evidence of rapid and accurate perspective-taking. I then suggest that the concept of joint attention may offer fruitful ground for future empirical research on ostension. I argue that ostensive communication is best conceived of as instances of joint attention, and that this model of ostension more successfully circumvents the criticism of cognitive implausibility than other minimalist accounts of ostension. Finally, I describe possible avenues for future research on ostension and joint attention.

Chapter 2

Mindreading in language evolution

In this chapter, I discuss the role of mindreading in evolutionary linguistics. Section 2.1 surveys the current literature on evolutionary linguistics, focusing on the progress that has been made by research focusing on the role of language learning and transmission between generations. This section was first published as "Models of Language Evolution", a peer-reviewed chapter by Cathleen O'Grady and Kenny Smith published in 2018 in the *Oxford Handbook of Psycholinguistics*, reproduced by permission of Oxford University Press (O'Grady and Smith 2018). This chapter was conceived and edited by both authors, and I carried out the research and writing. It is reproduced with the permission of Kenny Smith.

Section 2.2 discusses an important gap in this literature: how language users infer the meanings of signals. In much of the research discussed in Section 2.1, human participants (or computational agents) are supplied with signal-meaning mappings that they must learn and use; in other cases, they are not supplied with these mappings, but the inferences by which they establish them are not the focus of the research. In real-world communication, these inferences must be made constantly. I describe research on how signal-meaning mappings emerge in natural communication systems, and introduce the ostensive-inferential model of communication, which explains how these mappings can emerge simultaneously in human communication. I establish how ostensive communication underlies human language as a whole, and how this model of communication closes some of the gaps in the evolutionary linguistics literature by appealing to a central role for mindreading in language.

2.1 Models of language evolution

Abstract

This section reviews the models that provide evidence for the role of cultural evolution in the emergence of linguistic structure. This section discusses the levels of linguistic structure, and why the emergence of structure in language is a central question for evolutionary linguistics. It reviews the computational and experimental models which demonstrate that pressures operating during language learning and language use can give rise to the appearance of design in language, through the repeated cycle of learning and use that characterise language transmission. Finally, it discusses how learning biases at the individual level lead to the presence of typological universals: systematic patterns in how the world's languages tend to be structured.

2.1.1 Introduction

Language is startling in its complexity and expressive power. Unlike any other animal communication system, language provides a system for building complex signals from sub-components in a way that yields an endless set of possible combinations, capable of conveying an infinite array of possible meanings.

This open-ended expressivity arises through duality of patterning (Hockett 1960), which is the capacity of language to combine simpler structures to create more complex structures, at two distinct levels. At the first level, language combines meaningless units (that is, phonemes) into meaningful words and morphemes. For example, consider the English phonemes /d/, /b/, and /a/. None of these units have meaning on their own, but they can be combined into different meaningful combinations, such as "bad" and "dad". This is combinatoriality, and it gives us a highly efficient and expressive system. For example, most varieties of English have 40-45 phonemes (e.g. Ladefoged (2006) lists 44 phonemes in the inventory of Received Pronunciation), which can be recombined to form the 291,500 entries in the second edition of the Oxford English Dictionary (Oxford English Dictionary 2016).

At the second level of combination, meaningful words and morphemes are combined into larger, more complex meaningful units at the level of the phrase and sentence. For instance, the morpheme "bed" can be combined with the plural morpheme to create the word "beds", which in turn can be combined with other morphemes in a sentence like "The hotel room had two beds." This compositionality makes it easy for people to generalise rules they have already

learned to new items they encounter. For instance, young children might learn the word "bed" and its plural "beds", the word "dog" and its plural "dogs", and so on. Then, when presented with a nonsense word "wug", they generalise what they have learned about the plural morpheme to create the word "wugs" (Berko 1958). The same processes of generalisation apply at the level of multi-word utterances – compositional structure allows us to routinely produce and understand sentences that we have never seen (or that have never been uttered) before.

A major challenge for evolutionary linguists is to establish how a system like this can evolve, and why only humans have a communication system that works in this way. One potential explanation is that language is the result of cultural evolution (Christiansen and Chater 2008; Kirby 2001). Language, in common with many other human behaviours, is culturally transmitted: we are exposed to language by listening to those around us, and based on this linguistic input, we learn that language. Then, in turn, we produce linguistic output which forms the basis for language learning in others, who pass the language on themselves, and so on. This chapter will review a growing body of evidence suggesting that this type of multi-generational transmission process (sometimes known as iterated learning: Kirby and Hurford 2002) forces languages to adapt to constraints that affect how humans learn and transmit languages. Both computer models and experimental work on human learning suggest that this process of cultural evolution can result in the emergence of linguistic structure.

2.1.2 Computational models of language evolution

Early work investigating language evolution used computational modelling to simulate evolutionary processes acting on communication systems (Hurford 1989; Steels 1999). In a seminal paper, Kirby (2000) used computer simulations of iterated learning to demonstrate that compositionality could emerge in a model where each generation of computational agents learned a communication system from linguistic data produced by a previous generation that had learned in the same way. Kirby initialised his simulations with holistic systems (where meanings are communicated by signals that have no internal structure); as this "language" was passed from generation to generation, with each generation of learners searching for generalisations in the data they were presented with, compositional structure gradually developed, until after many hundreds of generations an elegant, compositional language had formed, where signals are composed through the rule-governed recombination of meaningful sub-components.

An important finding of this modelling work is the role played by learning bottlenecks in iterated learning systems (Brighton et al. 2005; Kirby 2000; Kirby 2002; Zuidema 2003). These

bottlenecks appear between generations, as one generation of language users presents the next generation of users with only a limited set of linguistic data, on the basis of which they are required to learn an open-ended expressive system. As languages are repeatedly transmitted through a series of bottlenecks, compositional structure develops.

The role of the bottleneck in driving the emergence of structure can be best illustrated through an example. Imagine a virtual world with moving shapes of various colours: there are squares, circles and triangles, which can be red, blue or green, and move in spirals, bounces, or straight lines. If each possible referent in this world is expressed with a holistic signal (that is, one signal for a blue square that bounces, another completely distinct signal for a blue square that spirals, yet another unrelated signal for a red square that spirals, etc.), 27 signals would be required in order to express the full range of referents. If a learner is only exposed to some of the 27 words (perhaps half of them), they have no way to accurately reconstruct the "correct" signals for referents they haven't seen labelled, because there is no pattern underlying how labels are associated with referents.

However, if the communication system is compositional, 27 meanings can be expressed using only nine signals (three signals expressing colour, three shape, and three expressing motion) and one rule for combining them (for example, combine them in the order colour-motion-shape). In this case, a learner does not need exposure to all 27 labels in order to learn the full language, but instead can infer the underlying system from a smaller number of observations (for an optimal generaliser, simply encountering each of the nine component parts once is sufficient), and subsequently exploit the compositional structure to produce labels which they didn't encounter while learning.

Whenever there is a learning bottleneck, language learners are forced to generalise, because the input they receive does not provide labels for all the meanings they might like to convey. This forces learners to hunt for commonalities across holistic labels (for instance observing that the syllable "ka" appears on two bouncing shapes, and then using it for all future bouncing shapes). Over time, these generalisations accumulate, and holistic languages gradually change to become compositional.

This seminal modelling work has been built on in two important directions. A related body of work focuses on phonological systems, showing that interaction and cultural transmission can explain how the phonemes found in the world's languages are organised in acoustic and articulatory space, and how combinatorial sound systems emerge (de Boer 2000; de Boer 2001; Oudeyer 2005b; Oudeyer 2005a; Wedel 2006; Wedel 2012; Zuidema and de Boer 2009). An-

other recent body of computational and mathematical modelling work has built on the earlier simulation-based results to extract general principles for how iterated learning and transmission bottlenecks shape linguistic systems (Griffiths and Kalish 2007; Kirby et al. 2007; Perfors and Navarro 2014), and to explore how cultural and biological evolution interact to shape languages and language learners (Smith and Kirby 2008; Thompson et al. 2016).

2.1.3 Language evolution in the lab

Computational and mathematical models continue to play a vital role in the emerging study of cultural evolution as it applies to language. However, scepticism regarding the applicability of these models to human language learners necessitates supplementing computational work with experiments using human participants. Experiments on cultural evolution are based on the same principal as models described above: they explore how data transmitted across generations of individuals is shaped by the process of transmission. In an experimental setting, this method of creating "transmission chains" of participants has been highly successful in understanding cultural evolutionary processes more generally (Whiten et al. 2016). In a standard transmission chain experiment, the first participant in a chain is presented with some material (e.g. a drawing to copy, or a miniature language to learn) and then required to reproduce it. This reproduction is then used as the training material for the next participant in the chain of transmission, and so on.

Importantly, each participant in a chain of transmission is not simply reproducing the material they are given, but rather forming a mental representation of it, and then using this mental representation to reproduce the material they were earlier presented with (Mesoudi and Whiten 2008). Because of this process of internalisation and recall, if participants have any pre-existing cognitive biases or expectations about the material, those biases will be imposed on their representation and subsequent recreation of the material, allowing for a transformation of the material as it is passed from person to person (Griffiths et al. 2008; Kalish et al. 2007).

The earliest study done using transmission chains, called "serial reproduction" (Bartlett 1932), found that the contents of certain genres of stories were transmitted more fully than others, and that story contents became distorted over time to match participants' pre-existing knowledge. More recent studies have used the method to study cultural change, finding that social information such as gossip is transmitted more accurately than non-social information (Mesoudi et al. 2006).

Cultural transmission therefore entails two vital steps: first learning the information, and

then reproducing it. An important question, then, is which of these two processes is responsible for the changes that occur to systems that are transmitted through iterated learning. Kirby et al. (2015) manipulated these pressures in a simple drawing task: participants in transmission chains were asked to look at a drawing created by the previous participant in the chain and reproduce it. In one set of chains, each participant was given time to memorise the drawing, and then reproduced it from memory; in the other set, participants were able to reproduce the drawing directly, while looking at it and without having to memorise it first. In the chains where participants had to memorise the drawing before reproducing it, the drawings became smaller and less complex over successive generations, and began to tend towards conventional cultural symbols such as numbers. In the chains that simply copied the drawing directly, there was no decrease in size or complexity, and no tendency towards symbolism. This suggests that learning, rather than reproducing, is what creates the bias towards compressibility.

These transmission chain methods therefore allow researchers to establish what kinds of information are best retained and transmitted, and infer what systemic biases might be at play during transmission (Mesoudi and Whiten 2008). Taken together, the growing list of biases can help to explain current cultural phenomena, such as religion, music – and language (Chater and Christiansen 2010).

Learnability and linguistic structure

Applying these techniques to artificial languages or communicative games allows us to investigate how communication systems evolve through cultural transmission, providing a close experimental analogue to the computational models reviewed in Section 2.1.2. In an early paper directly inspired by this modelling work, Kirby et al. (2008) taught participants a series of randomly-generated holistic labels for the set of 27 shapes described in Section 2.1.2 (27 shapes generated by combining three colours, three shapes, and three kinds of motion). After the training phase, participants were required to provide labels for these shapes, with their output used as the training input for the next participant in the chain. The transmission chain entailed a learning bottleneck between each generation: each participant was trained only on labels for a subset of the total set of shapes (14 of the 27), but was required to produce labels for the full set. This forced participants to provide labels for shapes they had never seen the label for, introducing a strong pressure for generalisation (although many did not realise that many of the test stimuli were previously unseen).

Although the labels provided to the first generation in each chain were random, the artificial

languages evolved over the transmission chains to become regular and generalisable. In an initial experiment, the labels became highly simplified (e.g. "poi" for anything moving in a spiral pattern, regardless of shapes or colour). Languages like this are highly learnable – there simply isn't much to learn – but not very useful for communicating, because each label would drastically underspecify its intended referent (e.g. "poi" would fail to distinguish between a spiralling black triangle and a spiralling red circle). In a second experiment, Kirby et al. (2008) manipulated each participant's output before passing it on to the subsequent learner in the chain, removing duplicate labels from each participant's test answers. This experimental manipulation was intended to mimic the natural pressure, acting on real languages, to be useful for communication. In this second experiment, the languages evolved over repeated episodes of transmission and developed compositional structure: sub-components of each label specified sub-components of the stimulus. For example, the first syllable might specify the colour, the second the shape, and the third syllable the kind of movement. This compositional structure makes the language simple enough for experimental participants to learn fairly accurately, while by-passing the ambiguity filter by providing every object with a unique label.

Subsequent work shows that a similar result is obtained if this artificial prohibition is replaced by actual communicative interaction. Kirby et al. (2015) ran an iterated learning experiment where pairs of participants were trained on an artificial language and then used it to communicate, taking turns to label pictures for each other – ambiguous labels would be problematic during this communicative task, but were not prohibited. The language produced during communication was then passed on to a fresh pair of participants, who in turn learned the language and used it to communicate, and so on. As expected, this repeated process of learning and use resulted in the gradual emergence of compositional structure. Kirby et al. (2015) also showed that the development of structure is dependent on transmission: when a single pair of participants play the same communication game over and over, compositional structure does not develop, showing that both learning and use play crucial roles.

Similar techniques have been used to explore the emergence of combinatoriality. As discussed in the introduction, combinatoriality allows us to use a few dozen speech sounds to generate tens of thousands of meaningful words. Hockett (1960) suggested that combinatoriality might therefore be a consequence of a pressure to create a large number of distinct meaningful signals – that is, as the potential number of meanings in a communication system grows, and it becomes more difficult to create new holistic signals, the communication system might begin to re-use components of the holistic signals in a combinatorial fashion. However, at least one

human language doesn't use combinatoriality to solve this problem. Al-Sayyid Bedouin Sign Language (ABSL), an emerging sign language, is still in the process of developing structure at the phonological level (which in the case of sign languages is achieved by handshapes that are meaningless on their own but combined into meaningful units) (Sandler et al. 2011). Despite lacking phonemes, ABSL is nonetheless capable of expressing a wide array of meanings, and is used for all the functions that other human languages are used for. This suggests that a fully-functional language can develop a large meaning space without combinatoriality, implying that a growing meaning space alone is not sufficient pressure for combinatoriality to emerge.

Verhoef et al. (2011) used an iterated language learning experiment to show instead that the pressure of learning holistic signals can, through cultural transmission, explain the emergence of combinatoriality. In an experiment in which participants are required to learn and then reproduce a sequence of 12 slide whistle sounds – which are holistic in the first generation of the chain, i.e. with no internal structure, simply random and continuous slide-whistle movements/sounds – they show that the whistle sounds begin to show structure, with internal components of the holistic signals beginning to be re-used within and between different sounds in the sequence. As a result, the sequence becomes more easily learnable and more accurately transmitted. This suggests that combinatorial structure can emerge independently of pressures arising from the number of meanings which need to be conveyed.

This finding is corroborated by Roberts and Galantucci (2012), who used a communication game to compare the effects of conventionalisation and number of meanings on the emergence of combinatoriality. Participant pairs, separated from each other at separate computers, were presented with a grid with simple animal drawings, and had to communicate to each other which drawing to select. Communication was possible using a stylus that distorted their drawings to prevent participants from simply sketching the animal to be communicated. Conventionalisation over the course of this repeated drawing game resulted in the emergence of drawings which exhibited combinatorial structure (sub-elements of drawings that were repeated across drawings), but the number of meanings was only weakly correlated with combinatoriality, which the authors suggest may be due to the limitations on the number of meanings that could be used in the experiment (in this case, only 20).

Properties of the communication medium may also contribute to the emergence of combinatoriality, namely rapid fade. Signals in spoken and signed language (and in Verhoef et al.'s slide whistle experiment) are transient: signals linger for only a short period of time (Hockett 1960). Galantucci et al. (2010), using a graphical communication task similar to that employed

by Roberts and Galantucci (2012), manipulated the rapidity of fade of signals by changing the speed with which communicative drawings faded from the screen. They found that more rapid fading led to a higher degree of combinatoriality. The number of meanings to be communicated was again not related to the degree of combinatoriality, providing further evidence against the hypothesis that larger meaning spaces drive the emergence of combinatoriality.

Communication and the emergence of iconicity

Transmission chain experiments such as these are therefore capable of explaining how linguistic structure arises. Perhaps surprisingly, similar methods have also provided insights into the very nature of linguistic signals, namely the fact that they are composed of arbitrary symbols for conveying concepts.

Garrod et al. (2007) use a Pictionary-like task involving a list of easily confused concepts – such as "art gallery", "museum", "parliament" and "theatre" – presented to a pair of participants. The Drawer in each turn must draw a randomly selected word from the list, while the Matcher must attempt to guess which of the words on the list the director is attempting to draw. As each pair of participants repeatedly plays this game, drawing the same concept multiple times, their drawings gradually become simpler and more abstract, and partners playing together converge, producing more similar drawings for a given concept. For instance, "cartoon" in one pair was initially drawn as a cartoon bunny and a bird, and over six rounds, simplified to become just a pair of stylised bunny ears. These results illustrate how iconic representations – representations that depict the content being communicated through resemblance – become abstract and symbolic through repeated use (Garrod et al. 2007), linked to their meaning only by convention.

Interaction appears to be an essential component of this process: transmission alone does not result in either convergence on shared representations or simplification (Garrod, Simon et al. 2010). However, in an experiment that combined the community-based game with a transmission chain, by periodically removing the most experienced member of the group and replacing them with a naïve participant, drawings became symbolic to the extent that newcomers were required to learn their community's conventionalised symbols for the list of concepts (Caldwell and Smith 2012). A similar effect is found when a drawing task occurs in a community-like setting that has eight participants interacting in a closed circle of shuffled pairs (Fay et al. 2010).

These graphical communication paradigms show how symbolic, arbitrary systems arise as a result of repeated interaction (whether in a pair or community), while the iterated learning paradigm shows how arbitrary systems become systematic through repeated learning. Theisen

et al. (2009) and Theisen-White et al. (2011) combine the two paradigms to demonstrate how the arbitrary symbols come to be used systematically, as they are in language. These studies use a graphical communication task with the potential for compositionality, by creating a list of items with shared semantic features: for example, five different kinds of entity (like people or buildings) through ten different themes (like education). Thus, the person in the education theme would be a teacher (Theisen et al. 2009).

As in previous experiments in the graphical communication paradigm, Theisen et al. (2009) found that drawings become increasingly arbitrary over time, while the arbitrary elements came to be used systematically: over repeated interactions, pairs of participants began to use increasingly symbolic elements to indicate components of the meanings they were communicating. For instance, if the first instance of "teacher" in a pair resulted in a drawing of a blackboard, subsequent school-related drawings ("teacher", but also "school" and "school bus") for that pair might include a simplified chalkboard element, plus a second simple component specifying which school-related concept was being conveyed. When a transmission chain was added, by using the first generation's drawings as training material for a second generation of interlocutors, the level of systematicity increased (Theisen-White et al. 2011). This suggests that both horizontal interaction and vertical transmission play a role in creating arbitrary and compositional systems, and that both communicative utility and learnability (by new players in the game) play an essential role in the emergence of structure, mirroring the results discussed above from Kirby et al. (2015).

Regularity and systematicity

The studies reviewed above focus on how languages and other communication systems are shaped by pressures inherent in their transmission – the requirements of learners to produce utterances for new meanings, or to learn and reproduce sets of rapidly-fading signals. A related strand of work (underpinned by a series of mathematical and computational models, e.g. Griffiths and Kalish 2007) explores how biases of learners, rather than external features of the transmission process, might also contribute to shaping language evolution. In particular, an intriguing body of work shows that even very weak biases in learning can have large effects on how languages are structured, because those weak biases accumulate over generations to create a substantial effect.

Some of the experimental work showing that this is the case has been concerned with the evolution of linguistic variation. Languages provide language users with multiple roughly equiv-

alent possibilities for particular forms, e.g. allophones (such as dark or light /l/), allomorphs (e.g. the past tense "-ed" is pronounced differently on the words "jumped" and "dragged"), or synonyms. The variant that is deployed in any given situation tends to be fairly predictable, being conditioned on sociolinguistic, phonological, semantic, or other criteria (Givón 1985).

Naively, we might therefore expect that the conditioned, predictable nature of variation in language therefore reflects a strong bias in language learning, strongly predisposing learners to condition or eliminate unpredictable variation wherever it occurs. Counterintuitively, adults appear to show no such tendency: if trained on unpredictable language data, they instead tend to probability match. For instance, if a particular variant appears 70% of the time in their training data, they tend towards using it at roughly the same rate and in a similarly unconditioned fashion (Hudson Kam and Newport 2005; Wonnacott and Newport 2005). This means that, in individual language learners, variant forms are preserved and remain unpredictable.

However, Reali and Griffiths (2009) show that this picture changes when unpredictably variable linguistic systems are passed along transmission chains. Using an approximation of synonymy in natural language, participants in their experiment were presented objects paired with labels. Each participant saw two labels for each object, with the two labels appearing with varying probability (e.g. for one object the two labels might appear in a 50–50 ratio, for another object the ratio of the labels might be 80–20). After training, when participants were asked to repeatedly label each object, single learners showed only a weak tendency towards regularisation, essentially matching the probability distribution of the input. However, across transmission chains, where these systems of object labelling were passed from person to person, there was a strong tendency towards regularisation, resulting in the loss of one of the variants.

One explanation for this result is that transmission chains automatically bring about the elimination of variation. However, Smith and Wonnacott (2010) find that variation can be maintained in a transmission chain, if that variation can become conditioned. In their experiment, adult learners learned and attempted to reproduce a variable system of plural markers – they learned a language in which the plural could be marked in one of two ways, with both forms occurring completely unpredictably. In line with the results from Reali and Griffiths (2009) individual learners did not exhibit a detectable tendency towards eliminating or conditioning this variation. However, after these miniature languages had been transmitted down chains, the variation was preserved (both plural markers lived on), but became predictable – each plural marker gradually became associated with a subset of the nouns, such that some nouns always took one plural marker and other nouns took the other. In their experiment conditioning on the noun

was the only possible way for the variation to become conditioned; in real languages, multiple such contexts exist, allowing for potentially complex but ultimately predictable systems of conditioned variation.

Collectively, these results provide reason for caution in extrapolating from individual-level experiments to assumptions about the emergence of linguistic features – finding no, or limited, evidence of a bias in a single generation of learners does not imply that the bias would not emerge as a factor on a population level.

Typological universals

Some features of linguistic structure are common to all (in the case of compositionality and systematicity), or virtually all (in the case of combinatoriality) of the world's languages. There are also many other features in which languages vary, for instance in the order in which words are typically combined. However, even here there are significant tendencies towards particular features, with multiple languages appearing to converge on the same structural solutions.

For instance, basic word order rules governing the sequence of subjects (S), objects (O) and verbs (V) in sentences could logically result in six different combinations (SOV, SVO, VSO, VOS, OSV, OVS). However, most of the world's languages use either SVO or SOV order. Although historical relationships between languages are likely to explain a number of these statistical tendencies (e.g. Dunn et al. 2011), a new and growing body of experimental work suggests that biases in learning are also likely to play a role in explaining such tendencies, in word order (Culbertson et al. 2012), morphological encoding of information (Fedzechkina et al. 2012) and phonological patterning (Wilson 2003).

Experiments using silent gesture paradigms are ideal for investigating questions like these, by allowing us to see which ordering strategies participants use when communicating in a novel medium. In these experiments, participants with no experience in any sign language are required to use silent gesture to express propositions, often as part of a communicative game, a little like the parlour game Charades. In these experiments, silent gesturers show a preference for SOV word order – for instance, when presented with a picture of a pirate throwing a guitar, a participant would gesture "pirate", "guitar" and "throw", in that order. This preference appears regardless of the dominant word order in participants' native languages (Goldin-Meadow et al. 2008).

However, although this might explain the prevalence of SOV in the world's languages, it cannot explain the prevalence of SVO. Schouwstra and de Swart (2014) show that the semantic

content of a message results in different word orders in silent gesture. Specifically, extensional verbs that describe the relationship between specific and existent entities, such as "throw" or "kick", are expressed with a preference for SOV word order. Intensional verbs like "imagine", which take objects that may be non-existent or non-specific, are expressed with a preference for SVO word order. Both preferences hold true regardless of whether a participant's native language uses SVO or SOV.

While these results go some way towards explaining why both SOV and SVO are common word orders in the world's languages, they are unable to explain why individual languages are usually consistent in their use of either SOV or SVO, rather than conditioning word order on the semantics of the verb. This tendency for consistency can be explained by cultural transmission. Schouwstra et al. (2016) combined the silent gesture task with a transmission chain approach, where groups of participants played a Charades-like communicative game while new group members gradually replaced more experienced individuals. While at the early stages each group's system of gestures showed a mix of SOV and SVO, conditioned on verb semantics, over time, word order became conventionalised to either all-SOV or all-SVO, as in earlier transmission chain experiments with artificial languages that found a tendency towards regularisation (Reali and Griffiths 2009; Smith and Wonnacott 2010).

More fine-grained typological patterns also appear to be mirrored in the biases of language learners. For instance, the majority of languages exhibit harmonic patterns governing the ordering of adjectives and numerals that modify nouns (Greenberg 1963): most of the world's languages have both post-nominal adjectives and numerals, or pre-nominal adjective and numerals (as in English: "blue cup", "three cups"), rather than a mix of pre- and post-nominal modifiers.

Culbertson et al. (2012) presented adult participants with an artificial language that had either mainly harmonic (adjectives and numerals both appeared pre- or post-nominally) or non-harmonic (adjectives appeared pre-nominally, numerals post-nominally, or the reverse) ordering, with a scattering of other word orders. Participants trained on languages which were mainly harmonic tended to make the language more harmonic, reducing the proportion of "noisy" other orders. However, participants trained on mainly non-harmonic patterns behaved differently, either simply maintaining a mix of orderings, or failing to reproduce the dominant non-harmonic patterns entirely. This suggests that the cross-linguistic preference for harmonic orders, as well as the scarcity of certain non-harmonic orders in the world's languages, may arise from learning biases on the part of language learners. Subsequent work has shown the presence of similar but

stronger biases in child learners (Culbertson and Newport 2015), and other typological patterns (Culbertson and Adger 2014).

Cross-species comparisons and biological evolution

Computational and experimental models of language evolution have shown how iterated learning is capable of generating compositionality, combinatoriality, and regularity in language; how communication interacts with cultural transmission to create an expressivity pressure that contributes to the emergence of structure; and how artificial language learning paradigms can illuminate the biases that are related to typological universals.

However, there are still a number of open questions in the field. Most notably, cultural transmission is not a process unique to humans: many species have culturally-transmitted repertoires of behaviour, and there are even other species who exhibit communicative behaviours that are culturally transmitted. Bird song, for example, is culturally transmitted in many species – birds learn their song early in development through exposure to species-typical song input. These systems have striking parallels with language: they have combinatorial (albeit not compositional or semantic) structure (Berwick et al. 2011); and although isolated songbirds do not acquire a fully-fledged system, but rather a degenerate and simplified one, this degenerate isolate song can revert to wild-type "natural" song when passed through a transmission chain of songbirds (Fehér et al. 2009). This suggests the existence of learning biases that shape the evolution of song (in zebra finches, and presumably in other species), much in the same way that human learning biases and cultural evolution shape the evolution of language (Claidière et al. 2014; Kirby et al. 2014).

These findings cohere with the models presented in this chapter, which predict the emergence of structure in an expressive system that is culturally transmitted (Kirby et al. 2015; Verhoef et al. 2011). Crucially, the nature of the expressivity pressure in birds is substantially different from that found in humans: whereas humans use signals to differentiate between possible referents in a communicative context, songbirds express mate quality through the size of their signal repertoire (Collins 2004). Similarly, many species have small repertoires of unlearned, holistic referential signals (for example, vervet monkeys have a small set of alarm calls signalling different kinds of predators) (Seyfarth et al. 1980), which is consistent with the prediction that communication systems emerging without a pressure for learnability will be holistic, with no compositional or combinatorial structure. Comparative studies investigating the similarities and differences between human communication and cultural transmission, and similar

systems found in other species, are likely to continue be a fruitful area of investigation for better understanding of our own cognition.

Finally, some accounts of the evolution of human language have argued that language is underpinned by language-specific cognitive apparatus, which has evolved through evolution by natural selection because the ability to communicate in this open-ended way is adaptive (e.g. Pinker and Bloom 1990). This suggests that linguistic structure is the result of biological adaptation, a reflection of a human-unique capacity for and predisposition to acquire communication systems with these properties. The evidence reviewed in this chapter suggests that cultural evolution is capable of explaining linguistic structure without appealing to additional, biological mechanisms: the processes involved in language learning and use are sufficient to result in the evolution of structure. But this is not to say that biological evolution should be ignored: a comprehensive account of language evolution will also need to explain the biological adaptations that led to our particular cultural environment, cognitive biases, and ability to learn complex signalling systems.

2.2 The mindreading gap

There is a crucial gap in this research on the emergence of linguistic structure: the question of how meanings become attached to signals. In much of the research discussed in Section 2.1, this link is pre-established. That is, in iterated learning experiments such as Kirby et al. (2008) and Verhoef et al. (2011), participants are provided with signals and their associated meanings during training, and the focus is on the effects of learning and transmission. In communication games (e.g. Roberts and Galantucci 2012; Garrod et al. 2007) and silent gesture experiments (e.g. Schouwstra and de Swart 2014), participants are presented with a constrained list of meanings, and are required to develop and interpret signals that refer only to this constrained meaning space.

In real-world communication, people are not given mappings between signals and meanings, or presented with a constrained range of meanings that can limit their interpretations of the signals they receive. In this unconstrained meaning space, people display an astonishing ability to infer the meanings of signals they have not encountered before, and to establish new signal-meaning mappings, in a range of contexts:

1. During language acquisition in infancy;
2. In emerging languages such as Nicaraguan Sign Language;

3. In non-linguistic communication, such as pantomiming, facial expressions, gestures, and communication that does not rely on the human body (such as signalling mechanisms on cars);
4. In linguistic communication, when resolving ambiguous or underspecified utterances (such as "Did you see that?")

This section explores the emergence of signal-meaning mappings and the role of mindreading in this emergence. In Section 2.2.1, I discuss the various routes by which new signal-meaning mappings can arise, including evidence that the simultaneous emergence of signal-meaning pairings is unique to human communication.

In Section 2.2.2, I explore how mindreading plays a role in the simultaneous emergence of novel signal-meaning pairings. Section 2.2.3 expands on this discussion to show how mindreading underlies not just novel signals, but linguistic communication in general. I discuss how an understanding of mindreading as central to language fills in some of the gaps left by research on language evolution.

2.2.1 The emergence of new signal-meaning mappings

A new communication system consists of new signal-meaning mappings. This is true of human-created systems (such as fictional languages, computer coding languages, or toy language systems in linguistic experiments), natural human linguistic systems (like emerging languages), and non-human communication (where "meaning" may in some cases be more accurately phrased as "response" – for instance, the response of bacteria to chemical signals in their environments). In human-created systems, there is no mystery about how meaning and signal become attached: someone creates the mapping and spells it out for others to learn. In natural communication systems, though, the emergence of these interdependent elements provides a puzzle. Do they evolve independently and become linked later? If so, what pressures lead them to evolve, and how do they become linked? Do they evolve simultaneously, and if so, what evolutionary scenarios could account for this?

An analytical model (Scott-Phillips et al. 2012) exploring the different routes by which communication systems may emerge shows that a state of non-communication is evolutionarily stable. This stability may be disrupted by two possible routes to communicative systems: essentially, the emergence of an action first, followed by a reaction; or the emergence of a reaction, followed by an action evolved to trigger the reaction. The first of these routes, *ritualisation*, is more common; this involves a behaviour that started out as a non-communicative cue and be-

comes communicative over time. For instance, a species that originally urinated out of fear at the boundaries of its territory would not initially be urinating as a communicative behaviour. The scent, however, would provide a cue to conspecifics: past this point is the territory of another individual. Over time, this behaviour may adapt to become a useful communicative signal. A less common route, *sensory manipulation*, involves the signaller manipulating the behaviour of a receiver directly, and this manipulative behaviour adapting to become a signal that induces a response. For example, a male insect that offers prey to a female in order to distract her and create an opportunity to mate is manipulating that female's behaviour; if the interaction adapts to become one of the female choosing to accept the prey as a signal of mating behaviour, it has become a signal-response pairing (Bradbury and Vehrencamp 2011; Maynard Smith and Harper 2003). Computational and mathematical models corroborate empirical work that suggests that these two possibilities account for the evolution of signal-response pairings (Scott-Phillips et al. 2012; Blythe and Scott-Phillips 2014). These routes may operate on individual, rather than evolutionary, timescales; for instance, the "nursing poke" observed in chimpanzees is an example of ritualisation based on an initially functional gesture (an attempt to move an arm to allow for nursing) that becomes a signal over time (Tomasello et al. 1994).

These models further suggest a third route – the simultaneous emergence of a signal and a response – that may occur in very specific conditions. Central to the simultaneous emergence of a signal and a response is the ability to "signal signalhood"; or, in other words, to make it clear that one's behaviour is communicative (Scott-Phillips et al. 2012; Blythe and Scott-Phillips 2014). As an illustration, consider how I might interpret a friend coughing: it may simply be a cough, but it may also be covert communication intended to signal something – scepticism, perhaps, or irritation.

In the absence of a pre-existing communication system, the fact that a behaviour is intended to communicate something is not a given; and yet humans can infer communicative intentions in even the most obscure behaviour, like a covert cough. The use of a yawn, eyebrow raise, or gaze direction can be used to convey a message – *I'd like you to leave now*; *Did you really mean that?*; *I can't speak openly but you should look at what I'm looking at*. This is possible even in communicative behaviours that do not use embodied signals like eye gaze or gesture: drivers may adapt signals externally visible on their cars to convey context-dependent messages to other drivers. For instance, a headlight flash may signal "there's a speed trap up ahead" or "you've forgotten to turn your lights on"; the use of hazard lights may signal "thank you" or "please be careful; there's something wrong with my car so my driving behaviour may be unexpected" –

or even, in very particular contexts, "caution: there are baboons in the road." Even in extremely constrained settings in an experimental communication game, participants are able to convey that their behaviour – restricted to moving a small avatar back and forth between squares on a grid – is communicative, to the point of being able to establish conventionalised codes using different movement sequences to communicate a range of different colours (Scott-Phillips et al. 2009).

Humans, then, are able to signal the signalhood of their behaviour. Signallers must be able to communicate that their behaviour is communicative, and design a signal that will be interpretable by their audience. Receivers must recognise that the signaller's behaviour is communicative, and establish the intended meaning of that communicative behaviour. This spontaneous creation of new signal-meaning mappings is what is seen in communication games and silent gesture experiments. Cultural transmission explains how linguistic signals, once established, become structured, as observed in emerging languages such as Nicaraguan Sign Language; the signalling of signalhood explains how signals are established rapidly and simultaneously in the first place.

A comprehensive understanding of the evolution of language must therefore include an understanding of how people are able to signal signalhood, recognise signalhood, design interpretable signals, and infer the meanings of signals. In the next section, I discuss these *communicative intentions* and the cognitive abilities that appear to underlie them.

2.2.2 Signalling signalhood: ostension and inference

When I signal signalhood, what I am doing is expressing my intention to communicate something. What exactly is meant by "communicative intention" differs between different accounts (see e.g. Sperber and Wilson 1986; Moore 2016a), but in common is their starting point with Grice (1957), who suggests that when a speaker "means something", this necessarily entails the speaker wanting the audience to recognise the intention they are expressing. Strawson (1964) emphasises that the listener's understanding must include recognition of the intention, otherwise communication has failed.

As an example, imagine I am trying to draw my toddler niece Lizzie's attention to the balloons attached to the fence across the street. I have an informative intention: I want Lizzie to see that there are balloons, because I think she'll enjoy looking at them. Imagine that Lizzie, however, is distracted and does not look up when I call her name. Then one of the balloons pops and startles her, causing her to look up and see the remaining balloons. Lizzie is now aware of

the balloons, but not because of my informative intention; she noticed the balloons on her own, and my informative intention has failed. If Lizzie had noticed me calling her name, looked up at me, and followed my point and description of the balloons to see them herself, my informative intention would have succeeded.

This requires not just an informative intention on the part of the speaker, but a recognition of this intention by the receiver. When this framework is spelled out in detail (as in Sperber 2000b; Scott-Phillips 2015; Sperber 2000a), it becomes complex, requiring each person to be able to think about the mental states of other people – for instance, I need to be able to recognise (i.e. mentally represent) Lizzie's lack of awareness of the balloons, and form an intention to change her mental state. More than this, I need to be able to represent not just Lizzie's mental states, but also mental states *that are themselves mental states* – that is, representations of representations, or metarepresentations. These representations are not necessarily conscious, but form a necessary part of the cognitive machinations that go into the informative intention.

This logic, played out, suggests that delivering, understanding, and confirming the success of an informative intention requires metarepresentation. In this example, each step will be indicated using square brackets and subscript numbers that indicate the levels of embedding involved in thinking about the proposition in question, with "0" being used for a mental representation of the world, and "1" being used to indicate first-order metarepresentation – a representation of a representation. Taking it step by step, we can see that delivering a Gricean informative intention requires me to entertain a representation of Lizzie's mental state:

- I am aware that $[_0$ there are balloons on the fence $_0]$ – My representation of a state of the world; not a metarepresentation since it does not include a mental state that is itself a representation.
- I am aware that $[_1$ Lizzie does not perceive that $[_0$ there are balloons on the fence $_0]$ $_1]$ – my metarepresentation of Lizzie's mental representation of a state of the world;
- I intend that $[_1$ Lizzie perceives that $[_0$ there are balloons on the fence. $_0]$ $_1]$ – another metarepresentation; that is, my intention to alter Lizzie's mental state by drawing her attention to the balloons.

Comprehending the informative intention would require Lizzie to entertain a second-order metarepresentation; that is, a representation of my mental state of intention, which is in turn a representation of her mental state of perception:

- I intend that $[_1$ Lizzie perceives that $[_0$ there are balloons on the fence $_0]$ $_1]$

- Lizzie is aware that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence $_0$] $_1$] $_2$]

If I were to introspect my own intention (thinking about wanting to tell Lizzie about the balloons), that would similarly require a second-order metarepresentation:

- I am aware that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence $_0$] $_1$] $_2$]

Most importantly, tracking the success of my informative intention – that is, was I successful in getting Lizzie to look at the balloons, or did she spot them for other reasons? – requires a third-order metarepresentation:

- I am aware that [$_3$ Lizzie is aware that that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence$_0$] $_1$] $_2$] $_3$]

This description does not account for how Lizzie knows that my behaviour is intended to inform her. The answer to this might seem obvious: I am using language, getting her attention, pointing, and meeting her eyes, all of which seem clearly involved in trying to tell her something. However, this is obvious to an adult human; we cannot assume that it is obvious to an infant who is the recipient of a communicative behaviour, and we certainly cannot assume that it is obvious in the case of a novel signal like flashing headlights or moving an on-screen avatar between squares in a communication game. Explaining how Lizzie understands that my behaviour is communicative – that is, the human capacity to understand that communicative behaviour is communicative – requires further explanation. This is best illustrated using an example that does not rely on established communicative channels like language or conventionalised gestures like pointing.

Imagine that my partner Johnny and I realise that we must post an important letter the next day, and I agree to do it in the morning. When the morning comes, however, I find I am running late and will not have time to post the letter. Rushing out of the house with no time to write a note, knowing that I will be in meetings all day with no time to text or call, and aware that Johnny will be home before me, I do the quickest thing I can think of and place the letter prominently on top of his computer keyboard on his desk, where I know he will see it soon after entering the house.

I do this knowing that he will understand the message: "I didn't get around to posting the letter; could you please take care of it?" That is, Johnny will understand that the placement of

the letter was a communicative behaviour, and will therefore be able to understand that I had an informative intention and work out what the message was by drawing on our shared context – that we'd discussed the previous night that the letter really needed to be posted today; that I'd agreed to do it in the morning, but clearly hadn't; that we'd not had time to talk all day; that I would have known that he would be home first, and known where he would be likely to go first in the house; and that we normally do not keep letters on his keyboard.

This probably sounds like an unruly amount of information to keep track of. That is a feature, not a bug, of this account, which argues precisely this point: that a communicative gesture like this entails a great deal of both context and mindreading from both the sender and receiver, and yet is somehow accomplished by people with something close to effortlessness. I would probably not think about it for longer than a couple of seconds if I had to come up with it while rushing to leave the house, and when I tested this hypothetical scenario on real-life Johnny, he decoded it instantly. This, despite the fact that a formal account of a *communicative intention* – that is, an action in which the signaller is attempting to make it clear that their behaviour itself is communicative – requires Johnny to entertain a fourth-order metarepresentation.

To recap, in order to interpret an informative intention, he will need to entertain the second-order metarepresentation that is my informative intention: Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$.

My communicative intention is a third-order metarepresentation: I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$.

This means that, in order to register my communicative intention – the signallhood of my signal, or my intention to make him aware that I am trying to communicate something – Johnny must entertain a fourth-order metarepresentation: Johnny understands that $[_4$ I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$ $_4]$.

If I am aware that Johnny got the message and my communicative intention has succeeded, I then entertain a fifth-order metarepresentation: I understand that $[_5$ Johnny understands that $[_4$ I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$ $_4]$ $_5]$.

On this account, the content of the communicative intention is the informative intention; and the content of the informative intention is the message. Inferring the content of the message from the communicative behaviour requires a certain degree of metarepresentation, while sending and tracking the communicative intention – a behaviour referred to as "ostension" –

requires an additional level of metarepresentation (Scott-Phillips 2015; Sperber 2000b).

An informative intention on its own is not sufficient to convey my message. If I had placed the letter somewhere prominent, but not unusual – for instance, on a hall table near the door – Johnny might have noticed it, but would not have considered my placement of it intentional; he might think I simply left the letter there to deal with it after work instead. He would not have seen the signal of the signalhood – the communicative intention that signals that there is an informative intention to perceive. It is *understanding that something is being communicated* that will motivate him to infer the content of the message (Csibra 2010; Sperber and Wilson 1986).

As I will discuss in Chapter 8, this particular account of ostension has been critiqued on the grounds of cognitive implausibility (Moore 2014; Moore 2016a; Geurts 2019). The point of dispute is not whether ostension requires mindreading, but rather what kinds of mental states this mindreading requires, and how many levels of metarepresentation are necessary. Even more minimalist accounts of ostension (Moore 2014; Moore 2016a) require mindreading.

Ostension allows any behaviour to become communicative: I can dance ostensively ("I particularly love this song!"), chew ostensively ("I'm trying to finish eating quickly so that I can respond to you"), or make eye contact ostensively ("The person talking has just said something that we were joking about together earlier.") In each case, there is no conventional meaning attached to the behaviour, but it is nonetheless communicative, allowing a meaning to be attached flexibly to the behaviour and interpreted in light of the immediate context.

An important point is that unsuccessful inference does not undermine the basic principles of this account. For instance, if I make slightly prolonged eye contact with a friend at a party in an attempt to communicate something discreetly, I will likely be referring to a recent or frequent topic of conversation between us. Perhaps someone is making a political point with which my friend and I both disagree, and this topic is something we discuss often and feel passionately about. I might intend my locking eyes with her and my stony expression to express disdain for the speaker's opinions.

My friend might misinterpret this signal. If she has tuned out of the conversation and is paying attention to the music, she might have noticed that the song currently playing is one that I had earlier mentioned disliking, and interpret the eye contact and stony expression to mean "Ugh, this song again." Her attention has settled on different features of our shared environment, and so her context for interpreting the ostensive cue is different from the context I had taken into account when delivering the message, leading her to an incorrect inference. She has still reached that meaning, however, by interpreting my behaviour as communicative, inferring that

I intend to inform her of something, and using contextual information to interpret the content of that informative intention.

Ostension is not limited to moments of communicative creativity using unconventional channels, like leaving a letter in an odd and conspicuous place. Rather, it has been argued to underlie everyday language use, because of the ambiguity inherent to linguistic expression (Sperber and Wilson 1986; Scott-Phillips 2015; Sperber and Origgi 2012). For instance, the words in the sentence "I'm hungry" mean that I would like to eat something, but the full meaning of the sentence is contextually dependent. If I said this while waiting for meals to be delivered in a restaurant, it might mean "I can't wait for our food to arrive"; but if I said it while recovering from food poisoning, it might mean "I think I'm finally on the mend." In the next section, I will discuss the role of ostension in everyday language.

2.2.3 Communicative intentions underlying everyday language

The *underdeterminacy* of a phrase like "I'm hungry" has been argued to apply not just to a subset of ambiguous utterances, but rather to be inherent to language, on the grounds that any utterance at all can have different meanings (wildly different or even just subtly different) depending on the manner of delivery, shared context, and speaker intentions (Carston 2002). While additional words can increase the precision of the idea that is communicated, "full explicitness (full encoding of content) is quite generally not achievable" (Carston 2013, page 24).

The incomplete encoding of thoughts into sentences is the result of incomplete encoding of concepts into words. This semantic underspecification is characteristic of natural language: nouns do not describe every aspect of their referents, specifying instead the relevant distinguishing factors (Carston 2013). For instance, nouns that refer to objects tend to specify shape, but not colour ("cat" refers to everything cat-shaped, regardless of colour), while nouns referring to material types specify substance, but not shape ("metal" refers to all things made of metal, regardless of shape). Nor do words generally encode a single meaning, signalling instead a range of possible meanings (Carston 2013). There is ambiguity in homophones like "bank" (a financial entity or the side of a river), as well as the ambiguity in polysemy that attaches multiple related meanings to the same word ("bank" in the financial sense may refer to the legal entity, the building with cash machines and tellers, the action of depositing a cheque, or a slang word meaning "lots of money"). We rely on context to establish which sense of the word to use; the sentence "He makes bank" coming from a native English speaker is probably using the slang meaning of "lots of money", rather than suggesting that someone is engaged in building a new bank.

Semantic underspecification is a feature of language that arises through cultural evolution, much like compositionality and combinatoriality. Languages in iterated learning experiments evolve to reflect features of the environment that are salient to interlocutors. For instance, Silvey et al. (2015) found that underspecification emerged in transmission chains when the salience of various distinctions was manipulated. By creating training materials for an artificial language that backgrounded a particular dimension (such as never requiring discrimination between shapes that share a kind of motion), the conditions were manipulated to create an expectation of salience for certain features, but not others. Across transmission chains, the emerging lexicons reflected the differing salience of the characteristics of the stimuli; for instance, developing consistent "morphemes" for shape and colour, but losing distinctive references for motion (Silvey et al. 2015).

Similarly, Winters et al. (2015) investigated the emergence of underspecification in response to changes in the situational context – that is, the information that was relevant to the message that needed to be communicated. Participants played a communication game that involved learning an artificial language, and the features of the objects they identified to their partners were manipulated. Languages evolved to encode the features that were distinctive and useful to the game. The results of both Silvey et al. (2015) and Winters et al. (2015) suggest that languages evolve to encode meanings that are salient given the situational context. This is consistent with accounts proposing that languages adapt to their socio-cultural environments (Dingemanse et al. 2013; Lupyan and Dale 2010; Perfors and Navarro 2014; Ramscar et al. 2010; Wray and Grace 2007), rather than being subject purely to general learning biases (Culbertson and Adger 2014; Fedzechkina et al. 2012; Reali and Griffiths 2009) or historical contingency (Dunn et al. 2011).

Underspecification and underdetermination are therefore central properties of language, and must be accounted for in our understanding of language evolution. In a radically underdetermined system, linguistic meaning can never be a literal decoding of the semantic content of each morpheme in the utterance, because that semantic content can only be established by using contextual information to resolve the underspecification (Sperber and Wilson 1986). That is, *all* utterances require context in order to be interpreted. This is true even of highly detailed communication that conveys as much background as possible: if I were to give this thesis to a stranger in the street, they would be baffled by why I did that – that is, they would not understand *what I mean by doing this*, despite the thesis including thousands of words of exposition establishing the context of what I am trying to communicate. When I give it to my supervisors, Postgraduate Research Office, or even family members, they have the context to understand why

I am doing so, and are therefore able to interpret the (very long) utterance.

In this way, any utterance behaves in much the same way as a novel signal: it provides evidence for the meaning that the speaker wants to communicate. The meaning must be inferred, because it cannot be recovered using strict associations between signals and meanings. Language greatly enhances the precision of the evidence for the speaker's meaning: if I accompanied my surly glance at my friend with a mutter of "Can you believe him?" or "This song again!", the additional contextual information would make it far easier for her to correctly interpret my non-linguistic ostensive signal by achieving greater precision in the contextual information I am giving her. The combination of conventional associations and inference makes ostensive-inferential communication both powerful and precise.

In other words, language "is made possible by mechanisms of metapsychology and is made powerful by mechanisms of association" (Scott-Phillips 2015, page 64). We can still communicate through a raised eyebrow, conspicuously placed letter or exaggerated chewing when we need to, but given the option, language is the best and most effective choice – which is why we choose these more covert methods when we cannot speak freely, do not have time to send a text message, or have our mouths too full to speak.

Linguistic structure allows for the precise communication of an infinite array of ideas, and an understanding of the evolution of language must account for the emergence of this structure, but this explains only part of the picture. A fuller understanding must explain how meanings are inferred from imprecise and novel signals, and how we signal the signalhood of our communicative behaviour.

The *ostensive-inferential model* of communication argues that the human ability to deliver and interpret communicative intentions explains these phenomena (Scott-Phillips 2015). Because of the centrality of inference in human communication, ostensive-inferential communication has been argued to be an evolutionary (and developmental) precursor to language (Sperber and Origgi 2012; Scott-Phillips 2015). As outlined above, ostensive-inferential behaviour relies on metarepresentation. Mindreading, then, is crucial to our understanding of language evolution.

2.3 Chapter summary

Language allows for vast expressive power. Research on language evolution has made considerable progress in establishing the role that cultural transmission plays in the emergence of lin-

guistic structure, but this research leaves a substantial gap: explaining how signals and meanings become linked. In the natural codes found in non-human animal communication systems, these signal-meaning mappings may arise through ritualisation or sensory manipulation as a result of evolutionary adaptation. In humans, signal-meaning mappings appear to be able to arise simultaneously, as demonstrated by the emergence of new languages like Nicaraguan Sign Language, lab-based communication games, and the use of novel signals to generate unconventional meanings.

The ostensive-inferential model accounts for this simultaneous emergence of signals and meanings by appealing to an advanced, flexible and efficient mindreading ability in humans, underlying communicative behaviours ranging from subtle cues to everyday linguistic communication. Empirical evidence on mindreading, though, presents a mixed and complex picture of human mindreading ability, in both infancy and adulthood: certain evidence suggests that even infants are capable of certain kinds of mindreading, while other methods suggest that mindreading ability develops later in childhood, and that mindreading in adults is effortful and error-ridden. This mixed picture presents a problem for the ostensive-inferential model, which has been criticised on the grounds of implausibility (Geurts 2019; Moore 2016a). In Chapter 3, I review the empirical and theoretical mindreading literature, identifying the implications of this research for the ostensive-inferential model.

Chapter 3

Who can read minds?

In Chapter 2, I made the case for the important role of mindreading in language evolution, drawing on the ostensive-inferential account of communication (Scott-Phillips 2015). According to this account, delivering and interpreting communicative intentions is enabled by recursive mindreading; and efficient, flexible mindreading therefore underpins the evolution of language as well as everyday language use.

It is not clear, though, that people are capable of mindreading that is fast and flexible enough to make this constant deployment plausible. The empirical picture is complex and sometimes contradictory, suggesting a complex interplay between mindreading abilities and language, an inconsistent and puzzling picture of mindreading abilities in infancy, and ongoing debates regarding the efficiency of mindreading in adults.

This chapter gives an overview of the current empirical and theoretical work on mindreading. In Section 3.1 I begin by briefly discussing what constitutes "mindreading", and how I will use the term in this chapter. In Section 3.2 to Section 3.5 I then review empirical work on mindreading in various population groups: typically developing human infants; infants and adults with autism and atypical language exposure; people across cultures; non-human animals; and human adults.

Section 3.6 gives an overview of various theoretical approaches to mindreading that interpret and reconcile this evidence in different ways, identifying three relevant "axes" of disagreement within the mindreading literature: nativism vs constructivism; one-system vs two-systems; and mentalising vs submentalising. I argue that individual positions on mindreading are positioned on all three axes simultaneously, and may share a position on one axis but not on another; for instance, a constructivist mentalising vs a constructivist submentalising account.

Section 3.7 explains the relevance of these theoretical positions to the ostensive-inferential account. I argue that the ostensive-inferential account rests on a one-systems mentalising ac-

count of mindreading (but is agnostic on the nativism vs constructivism axis), and that investigating rapid and involuntary mindreading offers a useful method for testing the predictions of this account. I review the current literature on rapid and involuntary mindreading in adults, and introduce the Dot Perspective task, the experimental paradigm used in this thesis.

3.1 What is mindreading, and who can do it?

The use of the term "mindreading" discussed in Chapter 2 remained in uncontroversial territory regarding what falls under the widely accepted definition of the term. That is, communicative intentions all involve entertaining *propositional attitudes*, which are a kind of mental state that consists of an attitude towards a particular proposition. For instance, "I hope that it doesn't rain later" consists of my attitude (hope) about the proposition of it raining later – the proposition is the content of the attitude. "I believe that he will arrive tomorrow" consists of my attitude (belief) about the content of the proposition "he will arrive tomorrow." The metarepresentation "I intend that Lizzie notices that the balloons are on the fence" consists of both Lizzie's propositional attitude about the balloons (her attitude of "noticing" the content "the balloons are on the fence"), and my propositional attitude about Lizzie's propositional attitude.

In the philosophical literature, mindreading has typically been considered to be the ascription of propositional attitudes to others (Lavelle 2018), making delivery and interpretation of communicative intentions an unambiguous a case of mindreading. This does not quite capture the full territory of mindreading, though, since there are psychological states that are not propositional attitudes, like "Tim is sad" – I can think about Tim being sad, and how that might affect Tim's behaviour, my interpretation of what he says, and so on, without thinking anything about what Tim is sad *about* (Lavelle 2018).

A survey of the empirical literature on mindreading quickly demands that we stray from this safe territory, as we encounter behaviours that look a lot like they have something to do with mindreading, but are not (or not indisputably) ascription of psychological states. When an infant displays behaviour that appears to be an instance of mindreading, some theorists might consider the best explanation for this to be ascription of propositional attitudes; some might consider it to demonstrate attribution of psychological states, but not necessarily full-blown ascription of propositional attitudes; and some might argue that the behaviour is best explained without appealing to mindreading at all (Lavelle 2018).

Some of the evidence on mindreading suggests that certain kinds of mindreading are less

demanding than others – for instance, developed earlier in infancy, or used by non-human apes. At this end of the "ease" scale is *visual perspective-taking* – that is, being aware of what another individual can see. Within visual perspective-taking, two levels are distinguished: Level 1 perspective-taking involves determining only *whether or not* something is seen, while Level 2 perspective-taking involves determining *how* it is seen. That is, while Level 1 perspective-taking will allow me to judge whether or not Miranda sees the mug between us on the table between us, Level 2 perspective-taking allows me to judge whether or not she sees the mug's handle when it is facing towards me (Flavell 1977). At the other end of the "ease" scale is recursive mindreading, which involves chains of embedded propositional attitudes such as those discussed in Chapter 2 ("I intend that she believes that I know x").

Much of the empirical research on mindreading is an attempt to establish the suite of mindreading abilities that is available to different people, ages, and species; the extent of advanced mindreading abilities like recursive mindreading in adults; and the efficiency and accuracy with which people are able to mindread. Collectively, this evidence presents a complex picture of who is able to mindread, and to what extent.

In the following sections, I first review the literature on infant mindreading, discussing the primary task that has been used for decades to assess young children's mindreading ability, and critiques of this task. I then go on to survey the evidence on the co-development of language and mindreading, and discuss what the developmental evidence suggests about infant mindreading. Sections 3.4 to 3.5 survey empirical mindreading work across species and across human cultures, as well as mindreading in adults, before I discuss the theoretical implications of this work in Sections 3.6 and 3.7.

3.2 Mindreading in infants

3.2.1 The explicit and implicit false belief task

The classic test of mindreading ability in children is the false belief task (Wimmer and Perner 1983). In the original task, children watched a puppet, "Maxi", hide a chocolate bar. Maxi then left the scene, and his mother entered, moving his chocolate bar from the cupboard where Maxi had left it, to a different cupboard. When Maxi returns, children are asked where he will look for his chocolate. If the child answers correctly (that is, that Maxi will look for the chocolate where he left it, since he does not know that it has been moved), then this is taken as evidence that the child has an understanding that other individuals have their own representations of

the world, and that this representation differs from the child's own representation. The classic finding, which has been replicated frequently enough to be considered a "developmental dogma" (Rakoczy 2017), is that children only begin to pass this task sometime after the age of four years.

This kind of *unexpected change* task – sometimes referred to as the Sally-Anne task, after the puppets in a later version (Baron-Cohen et al. 1985) – is only one variant of the false belief task. A different version is the *unexpected contents* task, in which a box that looks like it contains one thing (Smarties) turns out to contain something unexpected (pencils), with a similar assessment that involves a child being required to explain what a naive character thinks is in the box (Perner et al. 1987). A third version is the *misleading appearance* task, which involves an object that looks like one thing (e.g. a rock) but is actually something else (e.g. a sponge) (Flavell et al. 1983).

The false belief task has been widely criticised, and whether it is an acceptable test of early mindreading ability is the subject of much debate. One critique is that the task requires a suite of abilities beyond just theory of mind, most importantly fairly advanced linguistic ability, in order to keep track of the sequence of events and understand the question being asked (Bloom and German 2000). Another challenge to the "developmental dogma" is that children much younger than four years old behave as if they already have some mindreading abilities, as evidenced by early diagnosis of autism spectrum disorder (ASD). ASD is generally understood to be associated with a lack of typically developing mindreading ability, and is already evident before the age of four years (Bloom and German 2000).

There is also evidence that other mindreading abilities are developed before the age of four years, and before the false belief task is passed successfully: a sequence of experimental tasks that require children to demonstrate understanding of another person's mental state that differs from their own, and that are successfully accomplished by children in a particular developmental sequence (Wellman and Liu 2004). First come *diverse desires* tasks, which involve a child choosing between two snacks – for instance, a carrot and a cookie – and being told that an adult prefers the other snack. Then the child is told that it is snacktime, and asked which snack the adult will choose. If the child correctly indicates the snack that they did not choose themselves, this indicates that they are able to understand that other people have desires that differ from their own. The next task in the developmental sequence is the *diverse beliefs* task, which is very similar in format, asking children to choose whether they think a cat is hiding in the garage or in the bushes, being told that a character has the opposite belief, and asked where she will look for her cat. Children who can correctly answer this question are taken to understand that people have beliefs that differ from their own.

The *knowledge access* task shows children what is in a box, introduces a new character, and tells the child that the character has never seen inside the box. Children are asked whether the character knows what is in the drawer, with a correct answer indicating that they are able to keep track of what another individual knows. The false belief task follows, after which, at age five, children pass the hidden emotions task, which requires describing how a character will feel after being made fun of by peers, even if they continue to smile (Wellman and Liu 2004).

This developmental sequence has been found to be robust across children in the USA and Canada, as well as Australia and Germany (Kristen et al. 2006; Peterson et al. 2005), and largely robust across more widely varying cultures, with one difference: children in China and Iran pass the knowledge-access task earlier than the diverse beliefs task, swapping around the second and third tasks in the sequence (Shahaeian et al. 2011; Duh et al. 2016; Wellman et al. 2006). The evidence that children younger than four demonstrate some mindreading abilities, and that some mindreading abilities develop only after the false belief task is passed, suggests that success on the false belief task should not be taken as the single assessment of children's mindreading ability.

An alternative line of critique of the false belief task argues that it does not successfully assess children's understanding of false beliefs at all, because of its heavy reliance on children's linguistic understanding and expression. A range of alternative methods have developed to assess mindreading ability in "implicit" or "spontaneous" tasks, as opposed to the "explicit" or "elicited" original false belief task. These implicit tasks use a range of techniques, including violation-of-expectation, anticipatory looking and behavioural paradigms.

The violation-of-expectation paradigm has been used to establish the emergence of early cognitive skills, such as infants' understanding of the physical world (Baillargeon 1994). This method relies on infants' tendency to look for longer at events that surprise them, which suggests that this event violated their expectations about the world, and in turn provides evidence of what those expectations were.

Onishi and Baillargeon (2005) used an adapted violation-of-expectation false belief task with 15-month-old infants. The infants watched an experimenter playing with a toy watermelon, and then placing the watermelon into one of two boxes. Much like the Sally-Anne task, the watermelon was then moved into the other box when the experimenter could not see the scene. This was the "false belief" condition; in the "true belief" condition, the experimenter watched the watermelon being moved. In both conditions, the experimenter would then reach into a box – either the box into which the watermelon had been moved, or the now-empty box. The in-

fants were reported to look at the scene longer when the experimenter did not reach for the box that matched her beliefs – that is, if she reached for the box with the watermelon when she had not seen it moved, or when she reached for the watermelon's old location when she *had* seen it moved (Onishi and Baillargeon 2005).

Evidence of false belief understanding has been found in even younger infants using the violation-of-expectation paradigm: Surian et al. (2007) found that 13-month-olds looked longer when a caterpillar approached a hiding place that did not match his false beliefs. Song et al. (2008) found that infants' looking time not only reflected an awareness of an agent's false beliefs, but also changed accordingly when those false beliefs were corrected. Similar techniques have been used with the unexpected contents (Song and Baillargeon 2008) and misleading appearance variants of the false belief task (Scott and Baillargeon 2009).

Another method is the "anticipatory looking" paradigm, which tracks where infants look first during a task, inferring from this what infants expect to happen next. Southgate et al. (2007) found that 25-month-old children anticipated where a character with a false belief would search when she entered a scene. One anticipatory looking task has even reported evidence of belief computation in seven-month-old infants (Kovacs et al. 2010).

Behavioural tasks rely not on tracking infants' observation of a scene, but rather on children's elicited behaviour. In "active helping" tasks, children are given the opportunity to assist an experimenter in a setting that demonstrates whether they have understood something about the experimenter's mental states (not necessarily always false beliefs). For instance, in one such task, infants of 12 and 18 months were given the opportunity to play with two novel toys. One of the two experimenters they were playing with then left the room, and while they were absent, a third novel toy was added. When the experimenter returned, they responded with excitement, and asked the child, "Can you give it to me?", the children were able to do this successfully, which suggests that they had successfully tracked which toy was new to the experimenter (Tomasello and Haberl 2003). Other "active helping" tasks have found that 18-month-old infants were able to help an adult appropriately depending on the adult's belief about whether or not there was a toy inside a box (Buttelmann et al. 2009); and to track what an adult believed about the contents of a box they were reaching for, and hand the adult a desired object based on that belief (Buttelmann et al. 2014).

Rubio-Fernández and Geurts (2013) adapted the original false belief task to operate more like an active helping task. This version of of the false belief task still elicits responses from infants, but changes how those responses are elicited. This "Duplo" task , which uses Duplo figurines

instead of puppets, follows a standard false-belief procedure with some important differences. Instead of removing the Duplo character from the scene while the item is moved, the Duplo character remains on the scene, with her back turned, avoiding the child having to keep track of the character's presence as well as her awareness. The child is invited to discuss the character's perspective when her back is turned, and the experimenter provides the correct answer if the child does not: "Can she see me from where she is? She surely can't see me!". The discussion of the character's perspective is repeated after the item has been moved: "Did she see that? No she didn't!". When the Duplo character turns around, the child is asked to take over control of the character, with the prompt "What happens next? What is she going to do now?" This task found much earlier success than the classic false belief tasks, with 80 percent of three-year-old children showing the Duplo character move to the location where she had left her item. This suggests that the children were able to successfully track the Duplo character's beliefs, even if they would not have been able to successfully articulate those beliefs in a standard false-belief task.

This body of evidence from implicit tasks has, in turn, been criticised, largely on the grounds of failed replications (Kulke et al. 2018; Kulke and Rakoczy 2018). For example, a replication of Kovacs et al. (2010) demonstrated that the results were due to an experimental artifact (Phillips et al. 2015). A replication of the anticipatory looking task failed to find an effect in children, although it did find an effect in adults (Schuwerk et al. 2018); and the initial violation-of-expectation task (Onishi and Baillargeon 2005) has similarly failed to replicate (Powell et al. 2018). A range of currently unpublished studies (Kulke and Rakoczy 2018) have also found non-replications of Buttelmann et al. (2009) and Rubio-Fernández and Geurts (2013). A large collaborative replication attempt aiming to determine which of these results are robust is currently underway (Frank et al. 2017).

Because of these unsuccessful replications, it is difficult to draw strong conclusions from this literature. It is unclear how robust this evidence of early false belief is; the pattern of replications does not suggest that there is *no* evidence of early false-belief understanding in implicit tasks, but rather that it is complex and inconsistent (Kulke et al. 2019).

A further complication arises from co-existing bodies of evidence that suggest that mindreading plays a role in language development, but also that language development plays a role in mindreading. The following two sections discuss this literature.

3.2.2 Mindreading is necessary for language acquistion

Implicit false belief tasks are not the only evidence for early mindreading development. Early language acquisition implies mindreading, because the process of word learning is very difficult to explain without appealing to mindreading: how is a child able to determine the referent of a word that an adult provides? A very lean account might suggest that associative mechanisms are able to take care of this – that children hear a particular word in a particular context, and across a range of contexts, frequently enough to determine the referent of the word – but evidence suggests that children observe adults' visual perspectives, including pointing and eye gaze, to determine the referent of a new word (Baldwin and Moses 2001; Tomasello 2000).

Yurovsky and Frank (2017) used eye-gaze tracking to establish where 12- to 42-month-old children looked when an experimenter labelled a new item. The infants watched the experimenter play with two toys (established in a pilot task to have equal visual salience) and look at one of the toys while providing it with a label. Infants of all ages followed the speaker's gaze to the relevant item, but this behaviour became more reliable with age, since younger infants sometimes continued to look at the speaker's face rather than following her gaze to the toy. Children's tendency to gaze-follow was related to their success at correctly choosing the toy that the speaker had labelled. This suggests that infants rely on gaze-following to attach signals to meanings.

A similar study established that children rely on not just the speaker's perspective, but their referential intent, to determine the correct label for a novel item (Tomasello and Barton 1994). In this study, children watch an experimenter as they search for a "toma", looking through multiple containers, and eventually exclaiming happily when they find the desired item. In comprehension tests, children show an understanding that the items the experimenter touched while saying the word "toma" were not in fact the toma, and that this label was attached to the item that the experimenter eventually found.

This result was later replicated in children of 18 months, and an adaptation of the task found that children attached a novel label to an intentional, rather than accidental, action (Akhtar and Tomasello 1996). These results suggest that children do not simply attach a label to the object or action to which their attention is directed while hearing that label; rather, they rely on the speaker's intentions to correctly determine the referent of the label.

It is possible that these two mechanisms appear at different stages of the developmental process: that simple associations based on visual salience allow the acquisition of some initial vocabulary, which bootstraps the acquisition of further vocabulary and syntax, and that by two

years old, children are able to understand the speaker's description of her intention (e.g. "I'm looking for a toma") both lexically and syntactically sufficiently to understand that the label applies to the specific item for which she is searching.

That is, this body of results does not necessarily imply that children are able to interpret adults' referential intentions in their very early language acquisition. Nonetheless, these studies do demonstrate use of adults' behavioural cues (including eye gaze and pointing) in vocabulary acquisition, suggesting that mindreading is necessary for language acquisition.

Mindreading and linguistic deficits in autism spectrum disorder

Autism spectrum disorder (ASD) offers further evidence of the importance of mindreading in language acquisition. ASD is characterised by, among other things, impaired mindreading ability (Baron-Cohen et al. 1985; Baron-Cohen 1995; Surian et al. 1996; Baron-Cohen 1988). ASD is also associated with delays in language development; children with ASD produce their first words at 38 months, on average, compared to the age of 8-14 months for neurotypical children (Eigsti et al. 2011; Howlin 2003); this delay is often one of the first symptoms leading to an ASD diagnosis (De Giacomo and Fombonne 1998). Around 25 percent of people with ASD never acquire functional language (Anderson et al. 2007; Pickles et al. 2014; Wodka et al. 2013) and even autistic individuals with highly-developed language abilities may have lifelong difficulties with certain linguistic concepts (Eigsti et al. 2011; Tager-Flusberg et al. 2013).

These early linguistic delays seem to be due at least in part to difficulty in processing the social cues discussed in the previous section. Autistic children have been found to attach a label to the toy they are looking at, rather than the toy an experimenter is looking at (Baron-Cohen et al. 1997). This may have been the result of children with ASD not following the gaze of the experimenter at all; or of following the experimenter's gaze, but not using this as a cue for correctly inferring the referent of the label (Baldwin and Moses 2001), but a later study suggests that the result arises from ASD children looking at the experimenter's face less than neurotypical children (Preissler and Carey 2005).

Many children with autism are nonetheless able to acquire language by using alternative word-learning strategies. Additional cues from an experimenter, like pointing to an item rather than simply looking at it, or actually touching the object, improved the performance of ASD children in word-learning, although they matched the performance of neurotypical children only when they found the target object interesting (Parish-Morris et al. 2007). This more limited range of strategies explains how ASD children are able to acquire language, albeit with a delay

compared to their neurotypical counterparts (Markman and Wachtel 1988; Preissler and Carey 2005; de Marchena et al. 2011).

There are certain facets of linguistic competence that pose a challenge even to autistic individuals with highly-developed linguistic ability; for instance, difficulty producing and understanding pronouns that encode perspective (like "I" and "you"); previous knowledge (deictics such as "that" and "he") (Hobson et al. 2010; Novogrodsky 2013); and mental states (e.g. "know" and "guess") (Kazak et al. 1997; Ziatas et al. 1998; Tager-Flusberg 1992; Kelley et al. 2006). Even high-functioning individuals who no longer meet the criteria for an ASD diagnosis as a result of early intervention may continue to struggle with pragmatics (Kelley et al. 2006).

This evidence suggests that typical language acquisition and use depends on typical mindreading ability, which appears to play a role from the very early stages of word learning. At the same time, a different body of evidence suggests that language is necessary to develop mindreading skills. Section 3.2.3 discusses evidence for this in cases of language deprivation, and in typically developing infants.

3.2.3 Language is necessary for mindreading

Deaf children who are born into signing families generally have typical development of mindreading skills, but deaf children born into non-signing families do not. This suggests that early language exposure plays a role in these children's ability to succeed in mindreading assessments (Peterson and Siegal 2000; Meristo et al. 2011; Schick et al. 2007; Pyers and Senghas 2009).

This appears to be the case even when the linguistic demands of mindreading assessments are minimised, as demonstrated by a task finding that deaf children of hearing parents underperformed hearing children of hearing parents (Schick et al. 2007). This underperformance was predicted by ability with sentential complements (for instance, "I think *that xyz*"; i.e., the linguistic structure needed to encode a propositional attitude).

Even in cases where hearing parents make an earnest attempt to learn sign language, they have been found to use less mental state vocabulary with their children. The use of mental state conversation among mothers predicted variance in the children's false belief task performance (Moeller and Schick 2006). This could be a result of hearing mothers' sign language proficiency following a step behind their children's development, with vocabulary being learned by the parents only when children begin to take the lead on these kinds of conversations. Longitudinal research would be valuable in establishing the direction of causality here, although research from hearing children suggests that maternal mental state talk is likely to predict later false belief

task performance among children (Taumoepeau and Ruffman 2006; Taumoepeau and Ruffman 2008).

Children with cochlear implants, who have limited hearing abilities and whose parents may be advised against providing them with access to sign language (Lyness et al. 2013), may also have limited access to mental state vocabulary. A study found that mothers of these children used vocabulary commonly used by hearing mothers with younger hearing children, focusing on desires rather than "mental states" like "think" and "know" (Taumoepeau and Ruffman 2006; Taumoepeau and Ruffman 2008). This suggests that the parents were adjusting their use of mental state vocabulary to match what they estimated the child's understanding to be.

Chronic language deprivation in childhood seems to create a permanent deficit in mindreading ability, although ongoing language exposure appears to provide incremental improvements even among adults. Despite decades of sign language use, even signers around the age of 40 who had had only late access to sign language underperformed on advanced mindreading tasks relative to hearing participants and native signers (O'Reilly et al. 2014).

As mentioned in Chapter 2, Nicaraguan Sign Language is an emerging sign language. It began with a cohort of deaf children at Nicaragua's first school for children with disabilities in the 1970s, and has been transmitted to successive cohorts of children entering the school (Pyers and Senghas 2009). Many of the children arrived at the school with no language exposure other than the rudimentary "homesign" systems used to communicate with their families.

Pyers and Senghas (2009) used a minimally linguistic false belief task to assess the mindreading ability of different cohorts of signers. This task used a sequence of picture cards, requiring the participants to choose which of two options completed the story; this was done in order to distinguish between the linguistic demands of the task and the participants' ability to predict the behaviour of the characters in the story. The first cohort of signers, tested at mean age 27, had limited use of mental state vocabulary and lower performance on the task than the second cohort, tested at mean age 17. Half of the first-cohort participants did not use any mental state vocabulary at all, and these same participants failed the false belief task. Upon retesting two years later, both the use of mental state vocabulary and performance on the false belief task had improved among the first cohort.

Although both groups had begun learning NSL at a similar age, the language had undergone substantial developments by the time the second cohort began to use it; this additional level of linguistic sophistication may explain why the second cohort outperformed the first, despite the first having an additional decade of social experience. During the interval between the two

assessments, the second cohort – newly graduated – began to socialise with the first at a deaf association, which increased linguistic contact between the cohorts. This may have provided the first cohort of signers with a new level of exposure to mental state vocabulary and conversation, increasing their performance in the false belief task by the time of the second assessment.

The deficit does not appear to extend to all assessments of mindreading: a group of home-signers in Nicaragua, who had never attended school and had therefore been linguistically isolated their whole lives, underperformed both age-matched NSL signers and a group of Spanish speakers who had had minimal education on an explicit false belief task. However, they showed no significant difference on a task measuring visual perspective-taking, suggesting that while visual perspective-taking can be developed through nonlinguistic social experience, passing the explicit false belief task depends on exposure to language (Gagne and Coppola 2017).

An important caveat in this evidence is that it primarily relies on false belief tasks, which (as discussed above in Section 3.2.1) have considerable limitations as an assessment of mindreading ability. It would be a fair characterisation of the evidence, though, to say that people with language deprivation show atypical performance on some elicited and verbally-expressed mindreading behaviours, indicating that language is necessary for the development of these skills. Evidence from studies of parental language use, and tasks that involve linguistic training for young children before mindreading assessment, corroborate the role of language in mindreading performance, as discussed in the next section.

The effect of parental input and linguistic training on mindreading assessments

Research on parents' use of mental state verbs finds that the more of these verbs parents use with their children, the better the children's performance on mindreading tasks (Slaughter and Peterson 2011). Longitudinal studies find that increased mental state vocabulary precedes enhanced mindreading performance (Taumoepeau and Ruffman 2006; Taumoepeau and Ruffman 2008). The use of verbs follows the same pattern as the developmental mindreading scale: parents in two studies in New Zealand began by discussing desires, and only later began to discuss thoughts and knowledge (Taumoepeau and Ruffman 2006; Taumoepeau and Ruffman 2008). As discussed above, this could be a result of parents following their children's developmental lead, and beginning to adopt certain mental state vocabulary as children begin to discuss it more (Moeller and Schick 2006).

Individual differences between children may also play a role. Eye gaze in infants of 10.5 months old has been found to predict their production of mental state verbs at 30 months, and

this in turn predicts their false belief task performance (Brooks and Meltzoff 2015). A meta-analysis of 104 studies of children's language use and false belief task performance found a moderate correlation, with early language abilities predicting false belief task performance (Astington and Baird 2005).

There are different explanations for how linguistic ability improves mindreading task performance. One possibility is that language provides labels for mental states, and this enables children to conceptualise those mental states and begin to use them in reasoning and decision-making (Olson 1988). Another explanation is that learning how to use complement clauses like "she thinks that *X*" can itself provide a child with the cognitive architecture for developing metarepresentations (de Villiers and Pyers 2002). Mental state talk may simply provide explicit evidence of other people's minds, which would otherwise have to be inferred by behaviour (Harris 1996).

A misleading object task compared these different explanations (Lohmann and Tomasello 2003). Children were given the opportunity to interact with a misleading object (a toy flower that was actually a pen) and asked about a puppet's beliefs about the item. Then, different experimental conditions used different training phases, before eventually children were asked again what they expected a new puppet to think about the item.

In the "no language" condition, the training phase consisted of minimal language use, for instance simply saying "Look!" when demonstrating the flower's use as a pen; that is, giving the children experience with the object, without any linguistic training. In the "discourse only" condition, the experimenter did not use any mental state verbs or complement constructions (e.g. "she thinks that") in discussing the object (for instance, simply saying "What is this?" instead of "What do you think this is?"). The "sentential complement" condition used mental state vocabulary and sentential complements while discussing the object, encouraging the children to use similar constructions, but without highlighting that the object was misleading; this focused on training with mental state vocabulary and complement structures, but not the item in question. The "full training" condition combined these strategies in an interaction that used mental state vocabulary about the flower-pen, with children asked what they thought the object was, being asked to recall their initial beliefs once they had discovered its true function, observing the puppet's discovery narrated by the experimenter, and finally asked what they expected the puppet's belief to be.

Children in the "no language" condition performed the same before and after training, but there were improvements in all three of the other conditions, with the greatest improvement in

the "full training" condition. This suggests that being provided with the syntactic structure and vocabulary to discuss mental states assisted the children's performance, even when that discussion had not focused on beliefs about the object in question.

Westra (2017c) argues that the pragmatics of elicited tasks create a range of possible meanings that young children must select between, and that the intended meaning is likely to be superseded by other possible meanings. On the false belief task, for instance, a "reality bias" makes the true location of the ball more prominent than a false belief about its location. Added to this is prompting from the experimenter to think about this true location; a likely desire to help the mistaken agent; a great deal of experience being taught things by adults and then being asked to demonstrate this knowledge; and a lack of experience hearing mental states being explicitly discussed. All of these factors combine to make the "cooperative" meaning the most salient one for a linguistically inexperienced child: they are more likely to interpret the question as a request to help the agent to find the item.

This explanation covers why linguistic ability and executive function both correlate with false belief task success. A diverse desires task – the earliest item on the mindreading scale – does not have these pragmatic snarls: even very young children have experience discussing people's desires, and the question "Will Sam choose carrots or cookies to eat?" does not invite a range of interpretations. Children's pragmatic and linguistic experience, on this argument, explain not just false belief task performance, but also the mindreading scale as a whole.

The evidence from autism, language deprivation, word learning and language training therefore suggests that language acquisition relies on mindreading, but mindreading relies on language. This contributes to a complex picture on the development of mindreading from the infant false belief literature. In Section 3.2.4, I briefly discuss the implications of this evidence for a general understanding of infant mindreading.

3.2.4 Summary: early mindreading

The range of implicit false belief tasks (including anticipatory looking, violation-of-expectation, and behavioural paradigms) has been taken to point to a "developmental gap" in mindreading (Lavelle 2018). During this gap, it appears that children demonstrate an understanding of other people's beliefs, but only later develop the ability to use this understanding to respond accurately to the crucial question in the traditional false belief task.

The inconsistent picture of replicability means that it is not currently clear how robust the evidence from implicit mindreading tasks is. However, there is other evidence suggesting that

young children have mindreading abilities before they are able to pass the false belief task: studies of word learning suggest that language acquisition relies on gaze following and an understanding of speakers' referential intent, while evidence from autism suggests that impaired mindreading abilities are linked to delays in language acquisition and ongoing difficulties with certain linguistic skills.

At the same time, the findings from language training, parental language use, and language deprivation collectively suggest the opposite: that linguistic experience improves performance on certain mindreading tasks. This may not be indicative of mindreading ability so much as a reflection on the tasks themselves, since this is precisely the critique levelled at explicit false belief tasks – that they do not test mindreading independently of other abilities, including executive control and language. The finding that children are able to pass adaptations that change the pragmatic demands of the task (Helming et al. 2016) combines with this evidence to suggest that fairly advanced linguistic knowledge is necessary to pass the standard version of the task, but that children seem to have some understanding of false beliefs before they reach this stage.

Brooks and Meltzoff (2015) provide a useful way to understand this complex empirical picture: the eye gaze following of infants less than a year old predicts their mental state vocabulary at 2.5 years, suggesting that early mindreading abilities like gaze following allow for early language acquisition. Those same infants' mental state vocabulary at 2.5 years predicts their later false belief task performance, suggesting that linguistic training is necessary to pass the task. This suggests a bootstrapping process in which infants use mindreading for language acquisition, but certain kinds of explicit articulation of an understanding of others' mental states relies on typical language acquisition and use.

There is substantial debate about whether the appearance of mindreading in infants truly constitutes mindreading, with different accounts presenting varying interpretations of this evidence. I will discuss this further in 3.6.

A critical weakness in this research is its limited subject pool. Across the psychological literature generally, the vast majority of results come from "WEIRD" (Western, Educated, Industrialised, Rich and Democratic) populations – a problem that has been shown to provide a skewed picture of the cognitive reality of our species by making results specific to this limited population appear universal (Apicella and Barrett 2016; Henrich et al. 2010). There is evidence that mindreading behaviour varies across cultures (Slaughter and Perez-Zapata 2014; Wellman and Liu 2004), but the extent of this variation, and its significance, are debated. The following section reviews this research and its implications.

3.3 Mindreading across cultures

Mindreading development is not cross-culturally universal. The tasks in the developmental mindreading scale (Wellman and Liu 2004) are passed in a different order by children in the United States and Australia compared to children in China (Wellman et al. 2006), with American and Australian children passing the diverse beliefs task earlier than the knowledge access task, and the reverse pattern in Chinese and Iranian children.

Evidence from a study on children in the UK and Hong Kong suggests that parental linguistic input plays a role in cross-cultural differences in explicit false belief task performance: parents in the UK were found to have a greater degree of "mind-mindedness" in conversation with their children compared to parents in Hong Kong, and the UK children performed better on explicit false belief tasks. Within each culture, parental mental state conversation correlated with children's task performance (Hughes et al. 2018).

This raises an intriguing possibility. There are some Pacific Island societies that assert that it is impossible to know the contents of another person's mind, a norm that ethnographers refer to as the doctrine of "opacity of mind" (Robbins and Rumsey 2008). In societies that adhere to the "opacity of mind" doctrine, speculating about the intentions of others results in serious social penalties. Robbins notes that this is not necessarily all that much of a deviation from Western cultures, in which people may ruminate on "how difficult it is to see into the hearts and minds of others" (Robbins and Rumsey 2008, p. 408). The important difference is that "opacity of mind" cultures consider even attempting to do this taboo, and for similar reasons, do not trust people's introspection about their own intentions and thoughts. For this reason, conversation about mental states is rare in "opacity of mind" cultures. This would predict that children in these cultures underperform children from very "mind-minded" cultures on mindreading assessments.

Another important difference may be in cultural norms about child-directed speech. While certain Western parenting styles result in a great deal of conversational engagement with infants, even mimicking turn-taking and discussing mental states before the infant is able to speak ("What do you want? You want your bottle? Yes you do!"), this is not culturally universal. On the island of Yasawa in Fiji, for example, parents do not typically engage in conversation with infants or young children, or discuss mental states with them (Barrett et al. 2013). Separate from how mind-minded a culture is, cultural norms about conversation with children may also determine those children's exposure to mental state vocabulary and syntax.

On implicit tasks, it appears as though cultural norms do not lead to cross-cultural differ-

ences: children from Yasawa, Ecuador, China and Illinois have been found to show no difference on spontaneous false belief tasks (Barrett et al. 2013), although limited sample sizes and a wide age range in these tasks require some caution in the strength of the conclusions that can be drawn from these results. These results tentatively cohere with the argument that, although these cultures do not engage in conversation about mental states, they nonetheless have no difference from more mind-minded cultures in their actual use of mindreading in social interaction. As anthropologist Rita Astuti notes in a discussion of mindreading abilities among adult NSL signers (Astuti 2015), "How do these individuals behave in their daily lives? How do they take care of their children? Do they ever tell them white lies? What do they do when their children tell them lies? How do they react if they find out that their spouse is cheating on them? What do they do if they want to cheat on their spouse?"

On explicit tasks, the picture is mixed. Children in the hunter-gatherer Baka population in Cameroon successfully passed a false belief task at age five (Avis and Harris 1991), and studies have found identical false belief task performance across children in Canada, India, Peru, Samoa and Thailand (Callaghan et al. 2005) and across substantially different cultures within Indonesia (Kuntoro et al. 2013).

However, later studies focusing on children from Samoa (raised in an "opacity of mind" culture) did not replicate the results of Callaghan et al. (2005): the majority of the children did not pass an explicit false belief task before the age of eight (Mayer and Träuble 2013). Evidence that the opacity of mind doctrine is related to this false belief task performance comes from a study of Pacific Island families living in New Zealand, with mothers with a stronger Pacific Island identity engaging in less mental state conversation with their children (Taumoepeau 2015).

Another explanation for these results is children's understanding of what they are being asked. Young children's experience of social interaction – particularly WEIRD, middle-class children – involves a considerable amount of explicit pedagogy. That is, they are frequently asked to demonstrate their understanding of the world, and shown events and objects that they are later asked to discuss (Westra 2017a). This is not necessarily universally the case. In this light, consider an anecdote from Astuti (2015), who was using a false belief task with Malagasy children, and who sent the four-year-old participant's brother off on an errand while moving the coin hidden under one coconut shell to a new hiding place under a second coconut shell. The four-year-old child failed the task, and when his older brother returned, Astuti asked the brother to look for the coin, in order to demonstrate his false belief to the young research participant. The brother went straight to the new hiding place: "As he grabbed the coin and ran off, he shouted:

"That's why you sent me to fetch the water!" If this older child had failed the task, Astuti points out, it may have been as a result of his thinking through *what the missing agent would believe about why they had been removed from the scene*.

A similar anecdote from Avis and Harris (1991) corroborates this point: in a false belief task involving a chef hiding mango kernels to cook later, a five-year-old Cameroonian child said that the chef would be happy when he looked in his hiding place and found no mango kernels – because he would get his kernels eventually. This answer did not indicate a failure to understand the perspective of the chef, but rather reflected the child's expectations about what the person who had moved the kernels intended to do with them; her assumption, no doubt based on her experience of the world, was that they would return the kernels to their rightful owner, and she assumed that the chef would have the same expectation (Lavelle 2018).

These anecdotes are admittedly just anecdotes, but they suggest an important framing of cross-cultural variance in mindreading research: children being assessed will be trying to puzzle through why they are there, what the experimenter wants them to do, and why they are being shown what they are shown and asked what they are asked. They might arrive at different interpretations of the task, and these interpretations are will undoubtedly be affected by their experience of the world, mediated by culture.

The likelihood of different cultural interpretations of the task is further demonstrated by (Naito and Koyama 2006), in an investigation of why Japanese children pass the false belief task later than children from other cultures. When these children were asked to explain their response that the mistaken agent looked in the chocolate bar's new hiding place first, they often responded that the actor had promised to retrieve his chocolate. The children appeared to prioritise the agent's stated intentions in predicting his behaviour over his mental states (Slaughter and Perez-Zapata 2014).

Although the picture on cross-cultural mindreading differences is somewhat mixed, it is largely coherent with the evidence discussed in Section 3.2: implicit tasks (tentatively) appear to show a different pattern of performance from explicit tasks; and mental state language experience (as well as possible culturally-specific interpretive factors) plays a role in explicit task performance.

Humans are not the only social species to adapt our behaviour based on the behaviour of other individuals around us. A number of non-human species have demonstrated a range of mindreading abilities (Krupenye and Call 2019). This comparative evidence allows an assessment of the behaviour that is consistent across humans and other species, and an isolation of the be-

haviours that are unique to humans. In the next section, I review the evidence on mindreading in non-human animals, and discuss how this coheres with the infant mindreading literature.

3.4 Mindreading across species

Research into animal mindreading is what instigated the scientific and philosophical interest in mindreading of recent decades (Lavelle 2018), with the question "Does the chimpanzee have a theory of mind?" in the 1970s (Premack and Woodruff 1978). Since then, a great deal of comparative research has established a suite of social cognitive abilities across species, focusing mainly on canines (dogs and wolves), non-human great apes (bonobos, chimpanzees, gorillas and orangutans) and corvids (crows, magpies, jays, and so on) (Krupenye and Call 2019).

This body of research is complex, given the vast array of mindreading abilities to be assessed, and the range of species in which to assess them. A review paper (Krupenye and Call 2019) points to a wide body of evidence suggesting that non-human great apes – which I will refer to in this section as just "apes" – can keep track of what others can see: they have been found to choose food rewards not seen by competitors, conceal food from competitors, follow the gaze of others even around barriers, and re-check gaze direction if the target of attention is not obvious.

Apes also behave as if they are sensitive to other individuals' goals, intentions, and perceptual states. Much like young children, apes have been found to respond differently to accidental and intentional actions, to help others to achieve their goals, complete the failed actions of others, and respond differently to uncooperative and cooperative but incapable individuals. Similar sensitivity to the goals and intentions of others has been found in monkeys (rhesus macaques and capuchin monkeys), as well as understanding of both visual and auditory perspectives in monkeys (rhesus macaques) and corvids. Dogs have been found to steal food when a human experimenter cannot see them, and to follow the pointing gestures of knowledgeable rather than ignorant humans to find a food reward, and to follow an ambiguous cue to choose a toy that an experimenter can see, rather than one she can't.

Corvids have been found to take into account diverse desires, with male Eurasian jays – responsible for providing food for females during mating season – choosing food to bring to the female that was different from food she had just eaten her fill of (Ostojić et al. 2013). This was the case even when the female's desires likely contradicted the male's own desires – if a male had eaten his fill of one food, and the female had eaten another, he would still choose to bring her the food that he no longer wanted to eat himself but that she would likely still want, although

performance on this task suggested that he was still biased to his own preference (Ostojić et al. 2014).

Apes fail an adapted version of the false belief task on behavioural measures, although eye-gaze tracking does reveal some awareness of false beliefs: in a false belief task in which chimpanzees, bonobos and orang-utans watched a video of an actor looking for another actor in a "King Kong" costume, the apes' anticipatory looks were directed more towards the location where the searching actor falsely believed King Kong to be, than towards the true location of King Kong (Krupenye et al. 2016). A behavioural task found that apes were able to help an experimenter based on their false belief (Buttelmann et al. 2017). False-belief task based on looking times found null results in rhesus macaques (Marticorena et al. 2011; Martin and Santos 2016), suggesting that while perspective-taking and sensitivity to goals, intentions and desires may be found across taxa, understanding of beliefs may be unique to apes (or, alternatively, that the methods used to assess understanding of beliefs are best suited to apes).

Chimpanzees appear to take their audiences into account when communicating. A study evoking alarm calls found that the chimpanzees persisted producing the call and observing conspecifics' reactions until everyone had reacted (Schel et al. 2013a); and that chimpanzees were more likely to produce food-invitation calls for good friends and for high-ranking individuals than others (Schel et al. 2013b).

Chimpanzees have been found to rely more on gestural communication when the recipient was facing them than when the subject was facing away from them, and have been found to move into a recipient's line of sight before beginning to gesture and even deliberately solicit the attention of the recipient (Gómez 1996; Hostetter et al. 2001; Kaminski et al. 2004; Leavens et al. 2004; Liebal et al. 2004; Povinelli et al. 1996; Povinelli et al. 2003; Tomasello et al. 1994). This behaviour, which has been argued to be drawing the receiver's attention to the sign, fits the minimalist criteria of ostensive communication put forward by Moore (2016b), but not the more stringent criteria supported by Scott-Phillips (2016).

A comparison of captive chimpanzees and bonobos reared in environments with and without a considerable amount of human communication and socialisation found that the enculturated non-human apes performed better on social cognition tasks than the other group (Russell et al. 2011). The enculturated group had been reared in research projects that had exposed them to high levels of human contact, language training, and frequent communicative games and tasks, while the standard-reared group had been reared in zoos and laboratories did not have this level of contact. The enculturated group matched or outperformed 2.5-year-old children on a range

of tasks, while the standard-reared apes matched or underperformed the children. A later study found that enculturated chimpanzees and bonobos outperformed a standard-reared group on comprehension of declarative pointing and vocalisations (Lyn et al. 2010).

The evidence from the comparative literature therefore coheres with the infant literature in suggesting that a wide range of mindreading abilities are not captured by the elicited false belief task. Corvids, canines and non-human primates all demonstrate degrees of mindreading ability, ranging from visual perspective-taking to understanding of diverse desires and even implicit false belief understanding, as assessed by eye gaze. As in the infant literature, there is debate about whether the evidence from the comparative literature indicates genuine mindreading or merely the appearance of mindreading; I will return to this in 3.6.

So far, this chapter has discussed mindreading mainly in individuals that do not have the fully-fledged mindreading ability we find in typical adult humans – that is, non-human animals, infants, and individuals with ASD or who have experienced language deprivation. Even cross-cultural research is primarily developmental. This is because, after children are about five years old, the standard suite of mindreading tests begin to produce ceiling effects in typical adults in Western societies; there is not sufficient cross-cultural evidence to determine whether this is universally true. Research on mindreading in people beyond five years old therefore tends to focus on specific instances of mindreading, such as special populations (including autism, ASD, and people with brain injuries; Apperly et al. 2009). Mindreading research on typical adults must find inventive ways to circumvent the ceiling effect. The next section reviews the literature on mindreading in adults.

3.5 Mindreading in adults

In typical adults, mindreading studies generally explore the limits of mindreading abilities: rapid and involuntary mindreading; the accuracy and ease with which mindreading is used in interactive tasks; and the extent of recursive mindreading ability. Rapid and involuntary mindreading is the subject of the empirical work in this thesis, and so I will discuss this literature and its relevance to the ostensive-inferential model in greater detail in Section 3.7.1. This section surveys research on interactive mindreading tasks and recursive mindreading.

Recursive mindreading has been assessed using the "Imposting Memory Task" (IMT), initially with the goal of exploring the relationship between mindreading and schizophrenia (Kinderman et al. 1998). This task involves stories with complex plots and convoluted chains of

metarepresentation (for instance, "Jenny thought that Emma believed that the boss knew that the chemist did not want Emma to work for him"). These stories were read out loud to participants, who were then given forced-choice questions that each presented a pair of written statements about the story. One statement described something in the story correctly, and the other statement was incorrect. The statements either described the metarepresentational chains in the plot, or were memory tests of details from the story that were not psychological states.

The maximum embedding in the metarepresentations was five levels; that is, the participant would need to understand propositions like $[_5$ A knows that $[_4$ B intends that $[_3$ C believes that $[_2$ D knows that $[_1$ A wants $[_0\ x\ _0]\ _1]\ _2]\ _3]\ _4]\ _5]$. Control questions included up to six related non-mental plot points. The results from this task suggested a cap on metarepresentations at four levels of embedding. Up to this point, error rates remained low, but increased dramatically at the fifth level of embedding. Error rates on control questions were lower at all levels.

Rutherford (2004) found no difference in either accuracy or reaction time on a task that had a maximum of three levels of embedding in each question. Ten- and eleven-year-old children did show a drop in performance at three levels of embedding, with a ceiling effect at zero levels of embedding dropping to 58% accuracy with three levels, and 49% with four. The effect of a cap beyond four levels of metarepresentation was found again in a study that found a correlation between social network size and performance on the IMT (Stiller and Dunbar 2007).

The IMT, however, has a range of considerable design flaws that make these results difficult to interpret. Control questions consisting of related plot points were not recursive, providing an inadequate control to the much more challenging recursive mindreading questions, and many could be answered sufficiently by simply spotting a detail in one of the forced-choice statements that had not appeared at all in the story (for instance, a statement describing someone as a policeman when there had been no policemen in the story). The highly recursive syntax used in mindreading questions, compared to the simple syntax used in control questions, created an additional level of challenge in the mindreading questions that was not present in the mindreading questions, as did ambiguity in some of the mindreading questions. Finally, the mindreading questions sometimes were not fully recursive – one might, for instance, involve depictions of four individuals' mental states, but as two separate first-order metarepresentations.

In an adapted version of this task that attempted to eliminate these flaws, O'Grady et al. (2015) found that adults were able to respond with a high level of accuracy up to seven levels of embedding. This was true of both mindreading questions and control questions that involved recursive concepts (such as location, as in "the book on the table in the corner of the room").

It was also true of both explicit and implicit task presentations: in the implicit presentation, participants watched actors act out the stories, and then chose between two videos, one of which matched the story and one of which had incorrect details, while in the explicit presentation, the story and questions were read out on video by an experimenter. Although accuracy remained high throughout, participants' confidence in their answers reduced with levels of embedding (on both control and mindreading questions), and participants replayed the videos a greater number of times at higher levels of embedding with explicit presentation, indicating that they found the task more difficult. A similar task found that six-year-olds could track four levels of embedded mental states presented in a story acted out by puppets (for instance, "the owl knows that the squirrel wants the duck to know that the frog loves the duck") (Helming et al. 2015).

This evidence suggests that adults, and even children, may in fact be very good at recursive mindreading. There is a difference, though, between an ability to use mindreading in responding to stories in this way, and ongoing sustained representation of another individual's mental states through an interaction. A different strand of research, on mindreading in interactive tasks, presents a different picture of adults' skills in mindreading, suggesting that adults make simple and unexpected mistakes that prioritise their own perspective, even in tasks that require just Level 1 visual perspective-taking rather than seemingly more demanding recursive mindreading. Keysar et al. (2003) argue that there is a "distinction between having a tool, and using the tool as part of one's routine operation" – that is, while adults may very well be able to answer questions about others' mental states in settings like the IMT, they do not necessarily use this accurately or reliably in everyday interactions, instead appearing to have trouble setting aside their own perspective.

Keysar et al. (2003), Keysar et al. (2000) and Keysar (2007) use the "director task" paradigm to assess this *egocentric bias*. This paradigm requires a participant to move objects around a grid of 16 cubbies based on instructions from the "director". The participant can see all the objects, but some are occluded from the director. For instance, an instruction might be "move the small candle one row up." The participant can see all 16 cubbies, but four are occluded from the director. The objects in these cubbies are arranged such that the director's instructions will involve one perspective ambiguity. For example, there might be three candles – one large, one medium, and one small – with only the large and medium candles visible to the director. This means that an instruction such as "move the small candle" means the small candle from the director's perspective, but not the smallest candle from the participant's perspective. In order to choose the correct one, the participant must limit her search to the objects in the director's perspective.

Keysar et al. (2000) used a task like this with 12 different grid layouts for each participant and director pair. In six of these layouts, a control condition placed an object that could not be a possible referent in the critical occluded slot, in order to account for the natural visual sweep of the grid. Eye gaze tracking suggested that, on hearing the ambiguous instruction, participants nonetheless considered the items visible only to themselves as possible referents. Participants, on hearing the ambiguous instruction, looked first to the occluded slot, and only then to the target object, looking at the target object 584 ms later than in the control condition. The time taken to reach final decision and reach for the object to move it was 1,449 ms longer than the control condition. Participants even reached for the occluded object 23% of the time, only correcting themselves in a quarter of these cases, and actually moving the incorrect object the rest of the time.

Similar results were obtained in a follow-up task that required the participants to set up the array, making it clearer to them what the director could and could not see (Keysar 2007); and in a task that involved a participant hiding an item in a bag without the knowledge of the director, but still considering it a possible referent (Keysar et al. 2003).

Keysar et al. (2000) acknowledge that these results, while suggesting that participants may have some difficulty overcoming their egocentric bias, do not necessarily suggest that everyday communication is riddled with egocentric errors. Rather, they suggest that speakers may sometimes fail to tailor an utterance to the perspective of their audience, resulting in miscommunication, and that egocentric bias can explain the communicative errors that *do* occur. Speakers may nonetheless maintain an (imperfect) awareness of each other's mental states, they suggest, in order to reduce the chance of egocentric errors, and allow for repair when errors occur.

Rubio-Fernández (2017) argues that this task is not a good test of mindreading, and instead that the high demands it makes on selective attention (that is, paying attention to some objects and not others) create a confound: is it the selective attention introducing delays and errors, or mindreading? Lin et al. (2010) find that performance on the task relies on attentional resources, but conclude on the basis of this that speakers do not use mindreading in the task, because mindreading is too effortful. Rubio-Fernández (2017) argues that these results might be better interpreted as evidence that the director task is an insufficiently naturalistic and demanding task. Most importantly, it prevents participants from using the notion that people know about more than what is in their immediate perspective – an assumption that is universal in human communication. Moreover, perfect performance on the task could be produced without using mindreading beyond an initial decision to ignore the occluded cubbies, making it a poor test of

mindreading ability.

A range of studies have found results challenging the evidence from the director task that perspective-taking is effortful, laborious and inaccurate. A broad range of psycholinguistic literature shows that perspective-taking is commonplace in language, finding that speakers take into account their audience's knowledge and attention in designing their utterances, building on a body of shared mutual knowledge ("common ground"), and that this common ground is most effective when it is established collaboratively (Brown-Schmidt 2009; Hanna et al. 2003; Heller et al. 2008; Hanna and Tanenhaus 2004; Heller et al. 2012; Kuhlen and Brennan 2010; Lockridge and Brennan 2002; Nadig and Sedivy 2002). Additionally, a substantial body of evidence finds that adults (and children) engage in rapid and sometimes involuntary level-1 and level-2 perspective-taking (Surtees and Apperly 2012; Zwickel 2009; Freundlieb et al. 2016; Samson et al. 2010; Qureshi et al. 2010). This work will be discussed in greater detail in Section 3.7, and in Chapters 4, 5 and 6.

In the next section, I synthesise the evidence on mindreading discussed above, describing a range of theoretical positions in the literature that attempt to explain the development and extent of mindreading abilities in humans and other animals.

3.6 Three axes of mindreading accounts

Sections 3.2 to 3.5 review the current empirical evidence on mindreading. Although there is considerable uncertainty in this literature, there are certain crucial points that theoretical accounts of mindreading must explain:

1. WEIRD children reliably pass the explicit false task around their fourth birthday.
2. Before this age, children acquire sophisticated language abilities, for which they seem to require mindreading; this suggests that they are able to use mindreading for language acquisition before being able to pass the false belief task.
3. Somewhat tentative evidence from implicit false belief tasks further suggests that children have mindreading ability, including understanding of false beliefs, before they are able to pass the explicit false belief task.
4. Children with ASD have impairments in mindreading ability that appear to lead to delayed or unsuccessful language acquisition, again suggesting that mindreading plays a crucial role in language acquisition.
5. On the other hand, language appears to play a role in mindreading: success in the false

belief task depends on linguistic and cultural experience.

6. Non-human animals are capable of mindreading behaviours including perspective-taking and comprehension of desires and intentions.
7. Adults and children can accurately comprehend storylines involving recursive mindreading, but this does not necessarily mean that mindreading ability is continuously and accurately deployed in second-by-second communicative interactions. Evidence on adults' successful use of perspective-taking in communication is mixed, with some research suggesting it is error-prone, and other research suggesting it is an essential component of linguistic competence.

There are various approaches to making sense of the empirical results outlined above. They may be divided into broad camps, differing along three axes, which I will describe as the nativism/constructivism, mentalising/submentalising, and one-system/two-systems axes. Accounts take up a position on all three axes simultaneously: for example, a mentalising rather than submentalising approach may be either nativist or constructivist; a one-systems approach may be either mentalising or submentalising; and a two-systems account may be entirely constructivist or both nativist and constructivist.

3.6.1 Nativism vs constructivism

Nativist accounts (e.g. Carruthers 2013; Csibra and Gergely 2011; Scott and Baillargeon 2017) argue that the best way to explain the empirical literature is to posit that humans have innate mindreading abilities – that is, not learned. This does not mean that these abilities are present at birth, independent of experience, or universal across all humans. For instance, consider the capacity to reproduce, which is innate, but is not present at birth; is dependent on receiving adequate nutrition and social contact; and is not present in people with a range of medical conditions. Rather, innateness means that given the right context, the capacity will develop in a predictable way, as a result of non-psychological processes (like hormonal or other biological changes) rather than by psychological processes (like the formation of associations or inferences) (Lavelle 2018).

Different nativist approaches consider different types of mindreading abilities to be innate. Apperly and Butterfill (2009) defend the position that some mindreading abilities are innate, but that these abilities do not involve attributing propositional attitudes; rather, they extend only to attributing a very limited and inflexible set of non-propositional psychological states. They argue that the early literature suggests that infants may be capable of some inflexible and limited mindreading abilities, but it is only with social and cultural learning that children begin to engage

in fully-fledged attribution of propositional attitudes, as demonstrated by explicit mindreading tasks. This account will be discussed more thoroughly in Section 3.6.2; here, it is sufficient to note that this account is, at least partially, a nativist approach. The *Infant Mindreading Hypothesis*, meanwhile, argues that that the literature provides evidence of an innate human ability to attribute propositional attitudes to others (Scott and Baillargeon 2017; Carruthers 2013; Scott et al. 2016; Carey 2009; Carruthers 2011).

Nativist accounts explain the appearance of mindreading abilities in infants and non-human animals, as well as the appearance of some culturally universal mindreading abilities (Barrett et al. 2013; Scott et al. 2016), by appealing to innate processing mechanisms and concepts. Cross-cultural differences in performance on mindreading tasks can be explained as the result of different environments, leading to different channelling of innate concepts and mechanisms:

> *"Specifically, a nativist can claim that infants are innately endowed with certain core concepts (perhaps desire, belief, pretense, happy, sad, see, and tell) and certain basic principles of attribution (such as "seeing leads to believing"). Thereafter novel concepts can be acquired, and new principles of attribution learned, relying both on individual experiences and cultural input. So from this perspective it isn't surprising that culture might make a difference, nor that training might help performance.*" – Westra and Carruthers (2017, p. 166)

Constructivist accounts defend the position that mindreading is learned; that is, developed through cultural and social experience of the world (Gopnik and Wellman 2012; Gopnik 1996; Wellman 2014). The constructivist approach points to evidence of the scaling of mindreading beliefs (Wellman and Liu 2004; Wellman 2014) as evidence that children use cultural, social and linguistic learning to bootstrap a framework for understanding the mind, with some concepts being more difficult to learn than others, and relying on previously developed concepts to be intact before moving on to others.

The evidence from deaf children and adults, as well as cross-cultural studies that find differences in mindreading abilities at different ages, provide further support for this account. As in nativist accounts, there are differences between constructivist accounts: some argue that mindreading results from a cognitive system dedicated to enabling rapid learning about this kind of evidence (Wellman 2014), while others argue that mindreading relies on general cognitive abilities like inductive reasoning and working memory (Heyes 2014b; Heyes and Frith 2014).

Nativism and constructivism represent something more like two ends of a continuum rather

than two neat categories of accounts of mindreading: the role of environment and experience is accounted for in both, as is the presence of some innate cognitive abilities (specialised or not) in developing mindreading ability. Two-systems accounts, discussed in the next section, suggest a way to marry the two ends of the continuum in a single account of mindreading.

3.6.2 One-system vs two-systems

Apperly and Butterfill (2009) suggest that the empirical literature is best accounted for by two separate mindreading systems: "System 1" is early-developing, resource-efficient, and limited to simple forms of mindreading; "System 2" develops later through learning, and is flexible, powerful, and less cognitively efficient. System 1 accounts for the efficient mindreading we see in everyday encounters like "has that pedestrian seen me heading towards them on my bike?"; while System 2 accounts for the vast array of mental state information that can be represented flexibly and recursively, as in the famous *Friends* line "They don't know we know they know we know!"

This proposal has an analogy in the development of number cognition. Very young children are able to process differences in small numbers of items – up to around four – and large ratio differences in large numbers of items. This simple, early-developing system is accompanied by a culturally learned set of skills in number reasoning, often deployed consciously, effortfully and inefficiently, but flexibly (Apperly 2011; Apperly and Butterfill 2009).

Similarly, a two-systems account of mindreading suggests that certain abilities (most notably Level 1 perspective-taking) are present from early infancy, and are deployed efficiently and automatically. System 1 achieves its efficiency partly through being informationally encapsulated – that is, inflexible, unable to incorporate contextual information, deployed only in a limited range of contexts, and characterised by "signature limits", or points at which the system simply is not available to solve the problem at hand. Apperly and Butterfill (2009) suggest that, for the sake of speed and efficiency, System 1 mindreading relies on simple relational rules and concepts like "if a conspecific's eyes are closed, she is not encountering any objects in the surrounding area" and "she has registered the watermelon in the yellow box, but not the watermelon in the green box." Concepts like "encountering" and "registering" are non-propositional versions of "seeing" and "believing" – that is, they do not involve a mental state that represents a particular proposition.

Because System 1 is limited to simple relational rules and is informationally encapsulated, another ability must then come into play when I read Othello and comprehend the complex fictional plot, which may be condensed to "[$_5$ Iago intends that [$_4$ Cassio believe that [$_3$ he intends

that [$_2$ Desdemona intend that [$_1$ Othello consider [$_0$ Cassio's rehabilitation $_0$] $_1$] $_2$] $_3$] $_4$] $_5$]" (Van Duijn 2010).

Informational encapsulation accounts for young children's inability to pass the explicit false belief task before a certain point, despite implicit measures suggesting that children process false beliefs and use them in their behaviour before this point. On a two-systems account, younger children fail this task because, while their System 1 is able to track that Sally has registered the chocolate bar in one box but not the other, this information is encapsulated and not accessible to the System 2 processes interpreting the experimenter's questions – that is, it is *subdoxastic*, or not available to conscious introspection and not integrated with reasoning and decision-making.

The alternative to a two-systems account is the view that there is a unified cognitive system that explains the evidence outlined above. One way to do this is to support a nativist account proposing innate, domain-specific cognitive modules, specialised to enable mindreading (e.g. Carruthers 2017); another is a constructivist approach positing mindreading abilities learned from very early in infancy, with linguistic experience explaining the gap between early mindreading abilities and later false belief task success (e.g. Wellman 2014). A different kind of "one-system" account is the submentalising approach.

3.6.3 Mentalising vs submentalising

The submentalising approach restricts the definition of mindreading to the kinds of abilities that make up System Two on a two-systems approach. That is, on this account, the "implicit" mindreading found in infants, non-humans, and automatic behaviour in adults is best explained not by mentalising behaviours, but by "submentalising" – that is, achieving similar ends to mindreading, but without ever representing a mental state (Heyes 2014a; Heyes 2014b; Heyes 2015; Heyes 2018).

Instead, the evidence for "implicit mindreading" is argued to be explained better by mechanisms used across multiple cognitive systems – learned associations and behavioural rules such as "if agent B is facing an item of food and has her eyes open, she is likely to take the food." This does not require representation of a mental state, but rather a probabilistic account of an event likely to occur based on a range of cues. Rather than visual perspective-taking, for instance – which entails representing another agent's visual experience – the same results could largely be achieved by "attentional orienting" – which entails looking in the direction that another agent faces.

The submentalising account is constructivist, arguing that mindreading is literally "mind

reading" and analogous to print reading, in that it is culturally learned (Heyes and Frith 2014). As with print reading, inherited developmental disorders may disrupt the process (dyslexia in the case of print reading, ASD in the case of mindreading); and as with print reading, cultures differ in the extent of their development and implementation of the practice. Children in cultures that emphasise mindreading are explicitly tutored in the subject, as they are with reading, eventually developing fluency. This does not mean that mindreading is purely a matter of training, and that a captive gibbon could be trained to engage in mindreading behaviour – not only does this training rely on language, but human-specific cognitive architecture is still argued to provide the "raw materials" for developmenting of mindreading skills, much as they provide the raw materials for print reading.

Heyes (2014b) identifies three different uses of the term "implicit mentalising": one, which she calls the "agnostic sense", considers behaviour to be "implicit mentalising" if it looks as though mentalising must be involved in the behaviour; Heyes gives the example of a dancer anticipating the moves of her partner as the "agnostic sense" of mentalising. The "contrastive sense" holds "implicit mentalising" to be behaviour that looks like mentalising, but is not; and the "assertive sense" considers behaviour to be implicit mentalising if it entails thinking about mental states but rapidly and automatically, rather than slowly and efficiently. Heyes proposes that the term should be limited to the assertive sense – that is, we refer to behaviour as implicit mentalising only if it genuinely entails thinking about mental states – and that the contrastive sense should instead be considered "submentalising". The agnostic sense, on the other hand, should remain genuinely agnostic: without evidence that a behaviour entails mindreading, the *appearance* of mindreading should not be considered sufficient.

3.6.4 Relationships between the axes

This discussion is clearly far from exhaustive, but sketches, in broad terms, some of the main lines of dispute in explaining the current state of evidence on mindreading. The axes are not intended to present an exhaustive review of dichotomous positions a mindreading account can take, but rather a useful way to think about some distinctions (which may operate more as continua than dichotomies) that are pertinent to the current mindreading literature.

As mentioned above, accounts of mindreading take up a position on all three of these axes (and axes of disagreement other than those identified here). The three axes are largely unrelated, in that the position on one generally does not dictate the position on another. There are some broad tendencies – for instance, current submentalising accounts are constructivist, but

a nativist submentalising account, describing mindreading-like behaviours as innate abilities of another type, is not a logical impossibility.

In the next section, I return to the ostensive-inferential account of mindreading introduced in Chapter 2, and connect it to the mindreading literature surveyed in this chapter. I argue that the ostensive-inferential account is a mentalising, one-systems account of mindreading that is agnostic on the nativism/constructivism debate. I then discuss rapid and involuntary mindreading as a useful avenue for assessing the ostensive-inferential account, and introduce a method for doing so.

3.7 Testing the ostensive-inferential account

Ostensive communication, as described in Chapter 2, relies on metarepresentation. That is, in order to deliver or interpret an ostensive signal, I must be able not just to represent the mental states of others, but do so recursively, which requires recursive sequences of propositional attitudes such as "Yolanda believes that it is raining" and "Christine intends that Yolanda believes that it is raining."

Where does the ostensive-inferential account fit into the system of three axes? The kind of mindreading required by ostension cannot be implicit mentalising in either the agnostic sense or contrastive sense identified by Heyes; it must be the assertive sense, and therefore genuine mentalising. The nature of the mindreading necessary for ostension also means that System 1 of a two-systems account is inadequate to the task. The metarepresentation underlying a communicative intention precludes the simple relational rules and concepts (such as "encountering" and "registering") that make up System 1 of a two-systems account. These concepts may suffice for the lowest level of the metarepresentation in a communicative intention (for instance, "Lizzie registers the balloons") but will not suffice for the higher orders, which rely on concepts such as intention. Further, the informational encapsulation of two-systems accounts presents a problem for ostension, which requires speakers and hearers to account for a wealth of contextual information in delivering and interpreting each utterance – information that would not be available to an encapsulated System 1. At the same time, ostension demands that metarepresentation is achieved rapidly and efficiently, and without significant cultural learning (given that it is taken to underlie language acquisition), which makes the flexible but inefficient and late-developing System 2 of a two-systems account similarly inadequate. While it is possible that an alternative two-systems account that aligns with ostension could be articulated, the prominent two-systems

account detailed by Apperly (2011) cannot account for ostension in either its System 1 or System 2.

Proponents of the ostensive-inferential model of communication generally support a nativist account that suggests that infants are able not only to attribute propositional attitudes to others, but also to entertain metarepresentations. They argue that, while theoretical description of metarepresentation may seem complex, this does not necessarily imply that the processing itself is difficult. As an analogy, observing the speed and angle of an oncoming vehicle to judge whether it is safe to cross the road entails a complex theoretical description, but the individual making the decision is capable of making this judgement without consciously representing algebraic equations (Scott-Phillips 2015). Sperber and Wilson (2002) suggest that human cognition is adapted to detect the relevance of incoming information, making metarepresentational communicative intentions easy for humans in the way that sonar is easy for bats. Csibra and Gergely (2009) propose a "natural pedagogy" module that allows infants to recognise communicative intentions by using ostensive cues, such as eye gaze and infant-directed speech. However, it is plausible that a constructivist model that accounts for early development of sufficient mindreading skills could equally well support an ostensive-inferential account of communication. The ostensive-inferential account need not necessarily commit to either a nativist or constructivist position.

The ostensive-inferential model therefore rests on a one-system, mentalising account of mindreading: it demands an early-developing, rapid, efficient and flexible system that represents not just others' mental states, but recursive chains of mental states. It is agnostic on the nativist/constructivist axis.

This observation suggests a useful avenue for assessing the ostensive-inferential account. Previous research investigating whether people are capable of the kind of recursive mindreading required by ostension has used adaptations of the Imposing Memory Task (IMT) discussed in Section 3.5, which tests comprehension of propositions like $[_5$ A knows that $[_4$ B intends that $[_3$ C believes that $[_2$ D knows that $[_1$ A wants $[_0\ x\ 0]\ 1]\ 2]\ 3]\ 4]\ 5]$ (O'Grady et al. 2015; Helming et al. 2015). Although these studies have found evidence of high-level recursive mindreading in both adults (O'Grady et al. 2015) and children (Helming et al. 2015), it is unclear whether these methods are an appropriate test of the recursive mindreading found in ostension specifically. Most notably, these tasks used stories showing a series of different characters with various mental states, whereas ostension instead relies on a back-and-forth reflection of mental states between two participants, reflecting only the mental states of the interlocutors. These methods

are also able to provide evidence only of accurate recursive mindreading, not necessarily the rapid and efficient recursive mindreading required by ostension.

An alternative possibility for testing the predictions of the ostensive-inferential account is to draw on empirical work investigating the predictions of two-systems and submentalising accounts of mindreading. A large body of empirical research is currently engaged in testing the predictions of different accounts of mindreading using tasks designed to elicit rapid and involuntary mindreading in adults. Evidence from these tasks can be used to discriminate between mentalising and submentalising explanations of the abilities tested, and to establish the degree to which this rapid mindreading is informationally encapsulated, as suggested by the two-systems account. These tasks may also be used to contribute to the ongoing debate about the plausibility of the ostensive-inferential model: although this work does not assess recursive mindreading directly in the manner of the IMT, a greater understanding of the rapid and efficient mindreading that is achievable in adults is informative for an overall picture of how mindreading can be used in everyday interaction. Ideally, these methods could also be extended to assess whether recursive perspective-taking appears to be similarly rapid and involuntary.

Rapid and involuntary perspective-taking therefore presents a useful technique to test the ostensive-inferential account. If the current evidence of rapid and efficient mindreading in adults is best explained by submentalising, then these tasks no longer contradict evidence of error-prone and effortful mindreading in communication (Keysar et al. 2003; Keysar et al. 2000; Keysar 2007), reducing the weight of evidence in favour of adults being able to utilise rapid and efficient mindreading in communication. If rapid and efficient mindreading is informationally encapsulated and inflexible, this would render the kind of contextual, background knowledge used in interpreting ambiguous or novel communicative signals inaccessible to this kind of mindreading, suggesting that it cannot be used in ostension. That is, evidence for either a submentalising or two-systems account of rapid and efficient mindreading is evidence against the plausibility of the ostensive-inferential model.

The following section reviews the evidence on efficient mindreading in adults and introduces the Dot Perspective Task, the paradigm used for the empirical work in this thesis.

3.7.1 Rapid and involuntary mindreading in adults

Research on efficient mindreading in adults often refers to the abilities in question as "spontaneous" or "automatic", which generally function as synonyms for "rapid and involuntary" – that is, the mindreading behaviour in question is engaged in without being necessary for the task,

likely without any intention to engage in that behaviour, and rapidly enough not to interfere in any meaningful way with task performance. Chapter 5 will discuss the concepts of automaticity and spontaneity in more detail and make more careful discriminations between their meanings, but as there is currently no general consensus in the literature about the definitions of these terms and any meaningful distinction between them, they are often used interchangeably. In this section, I will refer to the kinds of mindreading in question as "rapid and involuntary" rather than automatic or spontaneous, regardless of the terminology used in the research.

As with research on implicit mindreading in children, methods to demonstrate rapid and involuntary mindreading in adults avoid elicited tasks, relying instead on measures like reaction time and eye gaze. Senju et al. (2009) use the same kind of anticipatory looking task that has been used with infants in a false belief task comparing adults with ASD to a matched neurotypical control group. When the mistaken agent re-entered the scene, the neurotypical adults were more likely to look at the container where she believed her ball to be than at its new location. The same was not true of the ASD group. Both groups, however, performed the same on elicited false belief tasks, suggesting that while the ASD adults did not show the same involuntary processing of the mistaken agent's belief, they were able to reason through understanding of mental states to achieve the same level of competence as the neurotypical adults. This study, however, has failed to replicate (Kulke et al. 2019); and a comparable task has found the reverse result: eye-gaze tracking of participants watching a short cartoon of animate triangles found that ASD adults performed similarly to neurotypical controls on eye-gaze measures, but struggled to articulate the storylines verbally (Zwickel et al. 2011).

Adaptations of the anticipatory looking paradigm investigated whether adult participants tracked the belief state of the agent involuntarily. One study had three groups of participants: one was told to track the location of the ball, one the belief of the agent, and one was given no specific instructions (which made this condition identical to a standard implicit false belief task). The ball-tracking group showed anticipatory looking consistent with them tracking the belief state of the agent, even though this was not the goal of the task (Schneider et al. 2014); and a variant of this study demonstrated a similar effect through participants' movements of a computer mouse (van der Wel et al. 2014). These effects were dependent on overall cognitive load: when participants had to give a response about the location of the ball, they no longer appeared to unconsciously track the belief state of the mistaken agent (Schneider et al. 2014); and a task that required participants to track a stream of letters, remembering which letter had appeared earlier in the sequence, also found no involuntary belief tracking (Schneider et al. 2012).

A different paradigm shows participants a picture of a scene with a water bottle on the left-hand side of a table and a book on the right-hand side (Tversky and Hard 2009). One condition shows just the items on the table; another has a person sitting at the table, facing the participant (as if they were seated opposite each other at the table), with his hands in his lap; and in the third condition, the person is reaching for the book, which is on his left side, with his left hand. Participants are asked "In relation to the bottle, where is the book?" Participants describing the scene from their own perspective would be likely to describe the book as being to the right of the bottle; describing it from the other person's perspective, the book is to the left of the bottle. With no actor in the scene, most participants described the scene layout from their own perspective, with some using descriptions like "a foot away" or "across the table".

With the actor in the scene, around a quarter of participants used descriptions from the actor's perspective, with no statistically significant difference between the picture in which the agent simply sat in the scene and the one in which he reached for the book. That is, even though responding based on the actor's perspective was in no way necessary for completing the task, some participants nonetheless did so. This may be explained by participants intentionally responding from the perspective of the actor, or by the presence of the actor interfering with participants' intention to respond from their own perspective, and leading to errors. One possibility involves intentional perspective-taking, and the other unintentional perspective-taking, but both possibilities suggest that participants took the perspective of the actor even when it was unnecessary. Since there is cross-linguistic variation in how objects are referred to relative to other objects (Majid et al. 2004), cross-linguistic research on this task would be informative in determining how dependent it is on language-specific frames of reference.

Even the perception of agency seems to be sufficient to trigger perspective-taking, as demonstrated by a task using a film with animated triangles and a spatial location task with triangle distractors (Zwickel 2009). In these films, the self-propelled triangles are of different sizes and colours, and are all isosceles – i.e. with one shorter side at the "back" of the triangle and two equal, longer sides joining at the "front" of the triangle – which, together with the triangle's movement and "reactions" to what is in front of it, creates the impression of the triangle having a perspective. The triangles act out short interactions; for instance, in the film "Surprising", a small triangle knocks on a door and hides when the large triangle opens the door and looks out, then surprises the large triangle while its "back" is turned.

While watching these animations, participants were asked to respond to dots that would appear in the scene: if the dot appeared to the left of one of the on-screen triangles, the participant

should respond with a left keypress, and if to the right, with a right keypress. If the triangle was read as an agent with a perspective, including right- and left-hand sides, the triangle's orientation could be the same as the participant's, or different: if facing up, the triangle's left and right-hand side would be the same as the participant; if facing down, they would be opposite, as if the triangle were a person facing the participant. Participants responded to the dots more slowly if the triangle's "perspective" differed from their own when they had watched animations in which the triangle's mental states were necessary to explain the story, suggesting that perceiving the triangle as an agent was sufficient to trigger the participant to adopt the triangle's perspective even when it was irrelevant to the task.

The "Dot Perspective Task" (Samson et al. 2010) similarly provides evidence that adults adopt the perspective of an on-screen character when it is irrelevant to the task. The on-screen character in this case is a humanoid avatar, and participants are not coached to consider the avatar as an agent (unlike Zwickel 2009). This paradigm has generated a wide variety of adaptations and a great deal of debate about whether the results demonstrate mentalising or submentalising (Furlanetto et al. 2016; Conway et al. 2017; Cole et al. 2016; Michael et al. 2018); when the effect occurs (Bukowski et al. 2015; Gardner et al. 2018b); and who shows the effect (Drayton et al. 2018). The following section reviews the basic task and some early findings using this paradigm. Chapters 4, 5 and 6 then present a range of experiments using an adaptation of the DPT, along with reviews of the DPT literature that is most relevant to each study.

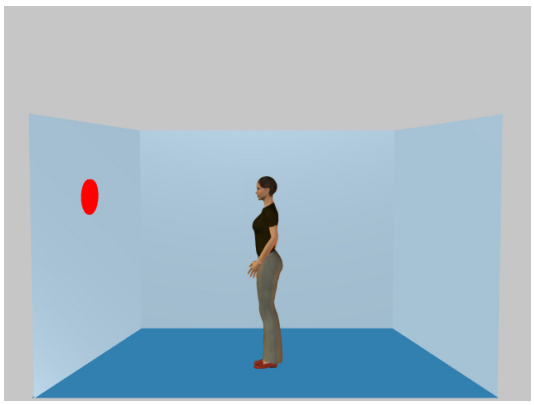
3.7.2 The Dot Perspective Task: evidence of efficient mindreading

In the original DPT (Samson et al. 2010), participants were shown a series of scenes each depicting a human avatar standing in the middle of a room, facing one side or the other (see Figure 3.1). An array of dots was displayed around the room, with different scenes showing different configurations of dots. In some scenes, all of the dots appeared in front of the avatar, making the avatar's perspective of the dots consistent with the participant's; for instance, there might be two dots on the front wall, meaning that the avatar and the participant both saw two dots. In other scenes, some of the dots were hidden behind the avatar, making the avatar's perspective inconsistent with the participant's; for instance, the avatar might only see one dot, while the participant could see two. In each trial, participants were shown a digit (e.g. "2"), followed by a scene, and asked to confirm whether the number of dots matched the pre-scene digit by responding "Yes" or "No".

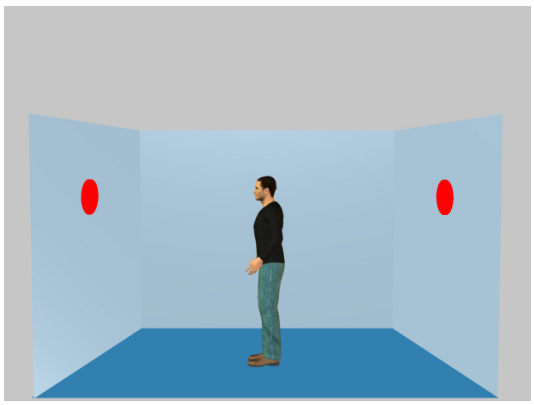
On some trials, participants were required to respond based on their own perspective ("Self"

trials), prompted by the word YOU appearing before the digit. On others, participants were required to respond based on the perspective of the avatar ("Other" trials), prompted by the word HE or SHE appearing before the digit. That is, a participant might be shown HE → 2 → and a scene in which the avatar sees two dots, requiring a "Yes" response. Response times were slower when there was inconsistency between the perspectives of the avatar and participant: if the participant could see dots that were hidden from the avatar, they took longer to confirm whether the pre-scene digit matched the scene.

This consistency effect appeared on Other trials, suggesting that judgements of what the avatar could see were slowed by the participant's own perspective, and suggesting *egocentric* interference – one's own perspective interfering with the calculation of another's perspective. Crucially, the effect also appeared on Self trials, suggesting that participants' judgements of their own perspective was slowed when the avatar had a different perspective. This *altercentric interference* suggests not only that participants were calculating the avatar's visual perspective rapidly, but also that they were unable to suppress their calculation of this perspective when it was irrelevant to the question being asked in a given trial – that is, the avatar's perspective was also being calculated involuntarily and unconsciously.



(a) A "consistent" trial, in which the avatar has the same perspective as the participant.



(b) An "inconsistent" trial, in which the avatar has a different perspective from the participant.

Figure 3.1: Stimuli from the original DPT (Samson et al. 2010). Each trial displays a picture of a room with an avatar and red dots appearing in front of and/or behind the avatar. Participants are required to judge how many dots they can see, or how many the avatar can see, as quickly as possible. Participants respond more slowly when the avatar's perspective is different from their own (*inconsistent* scenes, right) compared to scenes in which the avatar's perspective is the same as their own (*consistent* scenes, left). Crucially, this *consistency* effect occurs when participants are responding based on their own perspective (*altercentric interference*), suggesting that they are calculating the avatar's perspective rapidly and involuntarily even when it is irrelevant to the question.

Reasoning that switching rapidly between perspectives from trial to trial may have made

it more difficult for participants to suppress the avatar's perspective, Samson et al. (2010) conducted a second experiment in which Self and Other trials were each presented in separate blocks of trials, with two blocks of each presented in alternating order, and half of the participant group beginning with each type. This task found both egocentric and altercentric effects, and found that there was no significant difference in the size of the altercentric effect between this experiment and the task with mixed trial types. This suggests that being prepared to take the avatar's perspective at any time, or having taken it recently, was not necessary to produce the altercentric effect, contributing to the evidence that participants seemed unable to suppress the avatar's perspective even when it was task-irrelevant.

An alternative interpretation of this finding is that it was due to a spatial confound. That is, when the avatar's perspective is consistent with the participant's, all of the dots are arranged on one side of the screen. When the perspectives are inconsistent, the dots are arranged over a wider space, separated by the avatar. Either the wider spatial array or the effect of some dots being separated from others (or both) could increase the time taken to survey the scene, making it possible that this alone was responsible for the slower reaction times on inconsistent scenes. In order to control for this, Samson et al. (2010) conducted a third experiment that used a rectangular column, the same size as the avatars, as a control stimulus (see Figure 4.1). This was presented within-subjects, with each participant completing two blocks of avatar trials and two blocks of column trials.

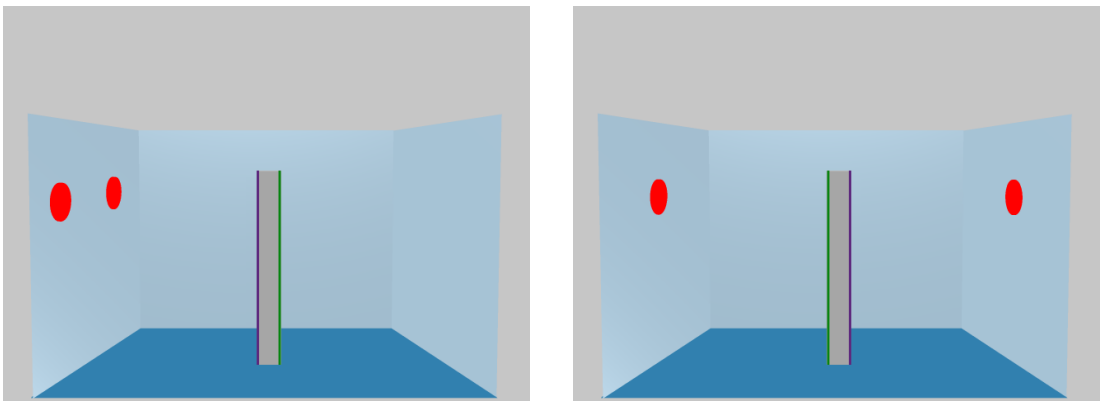


Figure 3.2: Stimuli from Experiment 3 of Samson et al. (2010), using a non-directional column as a control to determine whether the altercentric expect could be explained by a spatial confound.

A second adaptation in this task addressed a further possible criticism: that drawing attention to the avatar's perspective throughout the task is what drove participants to adopt that perspective, and that without this prompting, participants would not have trouble suppressing the avatar's perspective on inconsistent tasks. This task therefore did not ever ask participants to

take the avatar's perspective. Instead, they were told to ignore the stimulus in the centre of the room and take only their own perspective, an instruction that was reinforced by using the cue YOU before each trial instead of switching between YOU and HE/SHE. Despite this, there was an altercentric effect on avatar trials, but not on trials with the column, suggesting that taking the avatar's perspective was not necessary to induce the altercentric effect, and that a spatial confound could not explain the results.

The altercentric effect has has been reported in ten-year-old, eight-year-old and six-year-old children (Surtees and Apperly 2012). The effect also persisted when adult participants were required to complete a simultaneous distraction task that required them to respond to auditory stimuli with an incongruous key press – two key presses if they heard one tone, and one key press if they heard two tones (Qureshi et al. 2010). This task was used on two out of four blocks of DPT trials, allowing a comparison between the DPT results with and without the distraction. Overall reaction times were increased by the distraction task, but the altercentric effect remained, suggesting that despite the distraction, participants still tracked the avatar's perspective when it was task-irrelevant.

Qureshi et al. (2010) argue that this is evidence that Level 1 perspective-tracking does not rely on executive resources, and that it is perspective *selection*, rather than perspective *calculation*, that is disrupted by the executive function task. That is, they argue that perspective calculation happens automatically, and crucially, that its calculation is informationally encapsulated. Together, this evidence has been argued to provide support for the two-systems account: the level-1 perspective-taking demonstrated by the altercentric effect is rapid, involuntary, and does not rely on general cognitive resources.

Similar tasks have also suggested that this method demonstrates the presence of the signature limits predicted by the two-systems account, since this rapid and involuntary perspective-taking does not appear to extend to level-2 perspectives (Surtees et al. 2012; Surtees et al. 2013; Surtees et al. 2016b). A DPT adapted to level-2 perspectives used a cartoon avatar facing the participant over a table, with the numbers "6/9" or "8" displayed either on the table between the participant and the avatar (meaning that the avatar would have a different perspective on the number from the participant) or on the wall next to the table (meaning the avatar and participant would have the same perspective). This task therefore tests not just whether or not an avatar sees a dot, but *how* an item is seen. Although children between the ages of six and 11 were able to make judgements about what number the avatar would be seeing, there was no evidence of rapid and involuntary calculation of the avatar's perspective in the form of altercentric interference

(Surtees et al. 2012).

The presence of altercentric interference in level-1 but not level-2 judgements was confirmed in a later study (Surtees et al. 2016b). Level-2 perspective-taking appears to require individuals to mentally simulate the rotation of their bodies, which may be very demanding in certain contexts, and potentially explaining why it is not subject to the same kind of rapid and involuntary calculation as level-2 perspective-taking (Surtees et al. 2013). These findings again provide support for the two-systems account, with Level 1 perspective-taking mapping neatly onto System 1, and Level 2 perspective-taking appearing to be achieved using System 2. However, these interpretations have been widely challenged, with a range of alternative explanations for the data, and task adaptations exploring these explanations. The following three chapters will review these adaptations and explanations, presenting new evidence to clarify interpretations of the altercentric effect in the DPT.

3.8 Chapter summary

This chapter has presented an overview of the current mindreading literature. After briefly defining what "mindreading" entails, I reviewed the empirical literature on infants (3.2, discussing the range of tasks that have been used to assess mindreading in children, and the conclusions that can be drawn about the development of mindreading based on this evidence. In Section 3.3, I discussed the cross-cultural evidence suggesting that human mindreading development is not universal. Section 3.4 covered mindreading research in animals, showing that mindreading abilities are not unique to humans, but that other species have a more limited suite of abilities. In Section 3.5, I discussed current empirical mindreading research in adults.

Section 3.6 reviewed different theoretical approaches to making sense of the empirical literature on mindreading. I sketched three different "axes" of dispute within theoretical work on mindreading; namely, nativism/constructivism, one-system/two-systems, and mentalising/submentalising. I described how these different positions account for the current empirical evidence.

Finally, Section 3.7 argued that the ostensive-inferential account of communication relies on a particular account of mindreading: namely, a one-system, mentalising account (that is agnostic on the nativism/constructivism question). I suggested that empirical research on rapid and involuntary perspective-taking in adults presents a useful opportunity to test the predictions of this account, and reviewed current empirical work in this vein. I introduced the Dot Perspective

Task and survey early findings from this paradigm.

The following four chapters introduce a body of empirical work using the DPT, and commentary on this work:

- Chapter 4 surveys submentalising explanations for the altercentric effect, and presents an adapted DPT testing the mentalising vs submentalising accounts. This chapter also identifies and discusses a crucial inconsistency in the methods used between different DPT adaptations, explaining why this inconsistency has important consequences for the results and interpretations of these adaptations, and how this bears on the question of the "automaticity" of the altercentric effect.
- Chapter 5 presents a DPT adaptation that operationalises the methodological inconsistency identified in Chapter 4, and discusses the consequences of the results for both the submentalising and two-systems accounts.
- Chapter 6 explores an alternative explanation for the altercentric effect, finding null results in a task that is more complex than the standard DPT.
- In light of null results obtained in this and other DPT adaptations presented in this thesis, Chapter 7 surveys the literature on the replication crisis in the behavioural sciences, identifying the ways in which many of the problems identified as leading to a low replication rate are present in the DPT literature. Possible solutions for future research are identified.

In Chapter 8, I return to the ostensive-inferential account, discussing the implications of the conclusions of Chapter 4 to 7.

Chapter 4

The Dot Perspective Task: the effect of different stimulus types

Material in this chapter has been reproduced from the supplementary material included with O'Grady et al., submitted to the *Quarterly Journal of Experimental Psychology*. An early description of this research and initial analyses appeared in O'Grady et al. (2017). These papers were co-authored with Thom Scott-Phillips, Suilin Lavelle, and Kenny Smith, and all authors have given permission for the reproduction of this material here. The research was conceived by all authors, and all authors contributed to the writing and editing of the papers. I carried out the research, and analysed the data with assistance from Kenny Smith.

A submentalising interpretation of the Dot Perspective Task

As discussed in the previous chapter, the Dot Perspective Task (DPT) has been used to argue that people rapidly and involuntarily track the perspective of an on-screen avatar, resulting in a delay when the avatar has a perspective that conflicts with their own (Samson et al. 2010). This delay occurs when participants are asked to respond based on the avatar's perspective, suggesting *egocentric interference*. More crucially for the claim of involuntary perspective-taking, it also occurs when participants are asked to respond based on their own perspective, suggesting interference from the avatar's perspective, or *altercentric interference*.

The interpretation of this "altercentric effect" has been widely challenged (Santiesteban et al. 2014; Heyes 2014b; Conway et al. 2017; Cole et al. 2017). One critique has focused on the role of the central stimulus, suggesting that if a similar effect is produced with a non-human stimulus, this would support a submentalising interpretation of the results. That is, this would suggest that

participants were not representing the perspective of the central stimulus, but rather responding based on directional orienting: the avatar focuses the participant's attention in the direction that it faces, and this results in additional attention paid to dots in front of the avatar, with a delay in attention to dots behind the avatar.

Samson et al. (2010), Experiment 3, used a control stimulus to account for possible spatial confounds in the task, given that inconsistent scenes featured dots arrayed over a wider space, and separated by a vertical stimulus. This task found that there was no altercentric effect on a task that used a column as the central stimulus. This column, however, did not have the directional features of the avatar.

A range of studies have implemented various directional and non-directional controls in order to establish how much the specifically agentic, social, and directional features of an avatar explain the results. This chapter will survey these studies, identifying a crucial gap in the literature, and present a conceptual replication of the DPT that addresses this gap.

4.1 Is the altercentric effect specific to avatars?

There are various possible explanations for what features an avatar might have that could induce the altercentric effect. One possibility is that it is humanoid, which implies that cartoon characters of other animals would not have the same effect. Another option is that it is agentic, which suggests that a non-agentic object in the same position would not have the same effect, as confirmed by the column control in Samson et al. (2010), Experiment 3.

Santiesteban et al. (2014) argue that the crucial feature is that the avatar is directional. According to this "directional" hypothesis, the avatar functions as a kind of arrow, with the humanoid features that give it a front and a direction (eyes, forehead, nose etc.) indicating that one side of the room is of higher priority. According to the directional hypothesis, any directional stimulus should create the same effect, regardless of whether that stimulus has a "perspective" to be taken by participants. An altercentric effect[1] found with a non-avatar directional stimulus would suggest that this effect is not the result of perspective-taking (caused by holding two separate perspectives in mind simultaneously and suppressing one), but rather by directional orienting (being directed towards one side of the screen, and having to reconcile the number

[1]Note that the term "altercentric" strictly refers to the interference in one's own perspective caused by the perspective of another agent, in the same way that "egocentric" refers to interference taking another's perspective because of interference from one's own. The "altercentric effect" then could be argued to properly refer only to a result found as a result of perspective-taking. However, it will be used here to refer to the delayed response on Self-perspective trials caused by inconsistency in the array of dots, regardless of the cause of that delay.

of discs indicated by this directional cue with the total number of discs in the scene) – that is, a submentalising explanation may account for the altercentric effect better than a mentalising explanation.

The directional hypothesis was tested with an adaptation of the DPT using arrow stimuli (Santiesteban et al. 2014), closely matched to the avatars in colour and size, as controls that would not have a perspective to be taken. It is worth noting that, of course, cartoon avatars also do not have a mind to be represented by participants, nor a perspective to be taken. Avatars themselves are also not human, agentive, or animate. However, they do represent animate human agents with a perspective to be taken, and so will be referred to as having a perspective throughout this discussion. For ease of reference, other non-humanoid stimuli will also be referred to as having a perspective, which should be taken to mean the perspective that would be visible from the position of, and in the direction indicated by, that stimulus.

In Experiment 1, Santiesteban et al. (2014) asked participants to respond based on how many dots could be seen from their own perspective ("Self" trials, cued by YOU), from the avatar's perspective ("Other" trials, cued by HE or SHE), or based on how many dots the arrow pointed toward ("Other" trials, cued by ARROW). Self vs Other trials were presented within blocks, while Arrow vs Avatar trials were presented in separate blocks, with four consecutive blocks of each stimulus type, and the stimulus that was presented first counterbalanced across participants. The altercentric effect was found for both avatars and arrows, suggesting that participants experienced a delay on inconsistent trials whether or not the stimulus was humanoid. This effect was found in the first block of trials, suggesting that it was found in the arrows even before participants had experience with the avatar stimuli.

Switching back and forth between Self and Other trials in this way could conceivably result in participants maintaining an awareness of the perspective of the arrow or avatar throughout the task – that is, to use a perspective-taking-like strategy to more easily judge the number of dots to be seen from a particular vantage point in the cartoon room, and to have difficulty switching between this perspective and their own, leading to altercentric interference. Because of this possibility, Experiment 2 repeated the task with Self trials only. Participants were instructed to ignore the central stimulus, and arrow and avatar trials were mixed rather than presented in separate blocks. A significant altercentric effect was found for both arrows and avatars, which was interpreted as evidence that the altercentric effect is the result of mechanisms that are "not specific to the representation of mental states" – that is, to directional orienting rather than perspective-taking.

Both of these experiments have a potential flaw, in that there was a cue inviting the participant to consider the avatar as something with a perspective, either because they were explicitly asked to take its "perspective", or because it was presented alongside the avatar stimulus, inviting participants to consider the two stimuli as being similar. Although Experiment 1 found a consistency effect for arrows even before participants had seen avatars, this task required participants to explicitly consider the perspective of the arrow on half of all trials, potentially driving the consistency effect. In Experiment 2, although participants were told to ignore the arrow and avatar, trials with the two different stimuli were not separated into blocks. In this experiment, there was therefore no evidence that there would have been an altercentric effect if participants had completed only Self trials in a task with only arrows.

A between-subjects task using rectangular blocks as control stimuli reported that the greatest altercentric effect was found for avatars, followed by arrows, and then blocks (Nielsen et al. 2015). Task cues were modified so that the words YOU and HE appeared only in the avatar condition, to avoid participants being cued to think of as the other stimuli as somehow animate or social. The rectangular blocks had two different colours as vertical stripes, giving the blocks a "front" and "back" (see Figure 4.1); despite this, they are considered a "non-directional" stimulus since there is nothing inherent in the colours directing attention to one side of the block rather than the other, unlike an arrow or avatar stimulus that inherently has a front and back side. For the control conditions, the prompts were ROOM (indicating that the participant should respond based on the total number of dots in the room) and ARROW or GREEN (indicating the number of dots pointed to by the arrow, or by the green side of the rectangular block). The results suggest that the consistency effect is most substantial for the avatars, but that there is nonetheless an effect for both directional and non-directional non-humanoid stimuli.

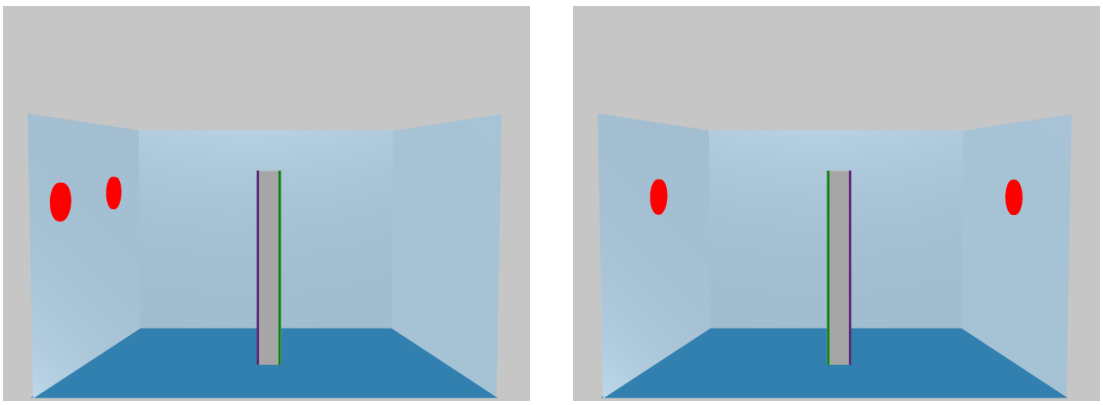


Figure 4.1: Stimuli from Experiment 3 of Samson et al. (2010), using a non-directional column as a control to determine whether the altercentric expect could be explained by a spatial confound.

However, the control stimuli in this task were not rigorously controlled for appearance. Arrows were small and black, positioned at the height of the larger, more colourful avatars' heads. Smaller, less attention-attracting stimuli may fail to produce an effect comparable to the avatars. The rectangular blocks were a similar size to arrows, striped half-green and half-black, with no clear front or back. Further, because participants were required to answer based on their own perspective as well as the perspective of the on-screen stimulus, the concern remains that participants are prompted to think of the object as having a perspective in order to more easily judge the dots visible from a certain point in the room. Additionally, a substantially different pre-trial procedure, involving repeated and longer displays of a fixation cross, may have played a role in causing participants to fixate on a certain part of the screen (see the discussion of fixation crosses, SOA and automaticity in Chapter 5). These differences may explain why an altercentric effect was found for the blocks, when previous work using a non-directional control stimulus in a task that required participants to respond based only on their own perspective throughout (Samson et al. 2010, Experiment 3) found no effect.

Considerable differences in the results of Nielsen et al. (2015) suggest some further substantive difference between this and other DPT variants. The effect sizes found in this study (a 162 ms altercentric effect for avatars, 116 ms for arrows, and 59 ms for blocks) are similarly not in keeping with other published effect sizes – an altercentric effect of 48 ms for avatars in the comparable in Samson et al. (2010), Experiment 1; and 60 ms across arrows and avatars, with no significant difference between them in the comparable Santiesteban et al. (2014), Experiment 1. Additionally, 13% of participants were excluded because of accuracy rates below 50%, in a task that standardly has mean error rates of less than 10% (e.g. mean 7.3% errors in inconsistent trials in Samson et al. (2010), Experiment 1).

Schurz et al. (2015) used a variety of inanimate stimuli in a DPT that gathered both behavioural and fMRI data with a range of control stimuli. Arrows were used as the "non-mental" control for avatars, while cartoon desk lamps were used for a stimulus with a direction but no conventional social meaning associated with that direction, unlike arrows. Finally, brick walls were used as a stimulus with no direction. In the first block of trials, participants were not asked about perspective at all, but instead asked how many discs appeared in the room, with the central stimulus varying between the four options. In the second block, participants were asked how many dots the lamp or arrows was pointing at, or how many objects the avatar could see. The wall stimulus was not used in this block of trials.

The behavioural data showed both altercentric and egocentric effects, with no interaction

between consistency and stimulus (arrow vs avatar; the authors report that the wall and lamp stimuli were not included in this analysis, but do not report why this was the case). Analysis of imaging results showed stronger activation in regions of the brain identified as relevant to mentalising (Schurz et al. 2015) during self-perspective judgements for avatars, but not arrows. Results also showed no difference in activation between arrows and brick walls, or arrows and lamps, in those regions where there was a difference between arrows and avatars. These results suggest automatic perspective-taking limited to humanoid stimuli, and not cued by directional stimuli. However, arrows and avatars differ in both appearance and animacy, meaning that any difference in neural activation may be a result of participants processing the visual appearance of a human body, rather than merely mentalising (Catmur et al. 2016).

A modified DPT that required participants to estimate larger number of dots by pointing to a position on a continuous number line that ran from 0 to 15 found an effect on error rates on inconsistent trials, with participants underestimating the number of dots when the avatar saw fewer dots than were in the scene overall (Marshall et al. 2018). There was no difference in this effect between arrows and avatars, which were presented within-subjects in a task requiring switching perspective between the participant and the avatar/arrow.

Two further tasks compared avatars to arrows, as well as to a camera, used as a non-social but directional stimulus (Wilson et al. 2017). This is of course not a social stimulus comparable to an arrow, but it should be noted that cameras are nonetheless instruments used from a human perspective, which may affect whether they are seen as having a "perspective". The black camera, positioned on a tripod, was the same height as the avatars; the red, mid-height arrow was intentionally designed to be visually easily distinguishable from the avatars, on the grounds that too much visual similarity could lead to confusion as to which stimulus was present in each very fast trial. Participants were cued by YOU to respond based on their own perspective, and by SHE (including the arrow and camera trials) to respond from the perspective of the stimulus, which was presented between subjects. There was no significant difference in altercentric effect between the three stimulus types. A repetition of this task manipulating the third-person perspective prompt, changed from SHE to FIGURE/CAMERA/ARROW, also found no effect of stimulus type on the magnitude of the altercentric interference effect.

While exploration of inanimate control stimuli in the DPT is widespread, there are still crucial gaps in the current literature. Namely, all of the tasks discussed above have used methods that could have induced an altercentric effect for arrows that might otherwise not have appeared: by presenting different stimulus types within-subjects; by requiring participants to take the per-

spective of the stimuli; or both. It is therefore not clear whether the altercentric effect would be found for arrows (or other inanimate stimuli) in a task that does not require participants to think about the number of dots the arrow points towards, and does not juxtapose the arrow stimulus with avatars throughout the task.

We therefore developed a modified version of the DPT that manipulated stimulus (arrows vs avatars) as a between-subjects variable. This task made no mention of the perspective of the avatar, or of the number of dots indicated by the arrow. Our hypothesis was that previous tasks have found an altercentric effect for arrows precisely because they presented the stimulus types within-subjects and/or required participants to take the perspective of the non-avatar stimuli throughout the task. We therefore predicted that we would find an altercentric effect for avatars, but not for arrows.

Although the principal hypothesis in this task was the question of altercentric effects in the arrow condition, the experiment also provided the opportunity to investigate an assumption in the design and analysis of my DPT variants. This assumption has resulted in large amounts of data being discarded, and may help with interpretation of why having conflicting perspectives results in a delay. We therefore also designed the task to investigate whether there was any difference in difficulty between trials requiring a "Yes" response, and trials requiring a "No" response.

4.2 Is rejecting the avatar's perspective harder than rejecting a non-perspective?

Participants in the DPT are required to judge whether the digit they are shown matches the number of dots in the given perspective; that is, to respond "Yes" on trials where the digit *matches* the number of dots, and "No" on trials where there is a *mismatch*. This means that the experiment design necessitates both *matching* and *mismatching* trials, to elicit both Yes and No responses.

Different kinds of *mismatching* scenes are possible in consistent and inconsistent trials. On consistent trials (where the avatar and participant see the same number of dots), any mismatching trial (that is, a trial with a digit that does not match the relevant perspective, eliciting a "No" response) represents a number of dots that is seen by neither the participant nor the avatar. It reflects nobody's perspective (the correct answer is "No, no-one sees that"). We call this trial type "No–none". On inconsistent trials, a mismatching digit may correspond to nobody's perspective ("No–none"), but unlike consistent trials, there is also the option for it to correspond to the irrelevant perspective (the correct answer is "No, I don't see that number, but the avatar

does"). We call this trial type "No–other". See Figure 4.4 for clarity.

Because of this imbalance, Samson et al. (2010) reasoned that *mismatching consistent* trials (all No–none) would be particularly easy, requiring the participant simply to note that the digit given did not match any agent's perspective; and that this particularly easy trial type was therefore not comparable to *mismatching inconsistent* trials (in this experiment design, all No–other), which should be more difficult. A number of researchers using the DPT have likewise assumed that No–none trials are easier than other trial types, and have therefore discarded all mismatching trials and analysed only matching trials (Qureshi et al. 2010; Samson et al. 2010; Bukowski et al. 2015; Cole et al. 2016). Not all DPT variants have discarded mismatching trials; Santiesteban et al. (2014) analysed mismatching trials in the first of their experiments, finding no difference between Yes and No trials (without further distinguishing between No–none and No–other).

The assumption that No–none trials are easier than other trial types has not been tested. This analysis has the potential not just to ensure that useful data is not discarded, but also to test the mechanism underlying the altercentric effect: perspective *calculation*, or perspective *selection*. Qureshi et al. (2010) argue that it is selection of the relevant perspective, rather than calculation, that results in the altercentric effect. If this is the case, this selection delay should similarly be found on No–other trials, which present a digit that matches the perspective of the avatar and should therefore require participants to disregard this perspective and select their own in order to correctly respond "No". By contrast, No–none trials (in which the digit reflects no agent's perspective) should not require this perspective selection, and should be faster than the other trial types.

An altercentric effect caused by perspective selection should therefore make the following predictions:

1. On inconsistent trials, No–none does not entail perspective selection, while both other options do, and so No–none should be faster than both Yes and No–other. This is the untested hypothesis that has caused most previous experiments to discard this data.
2. On consistent trials, neither Yes nor No–none requires perspective selection, so there should be no difference between these options.

If the data reflects these hypotheses, this would support the two-systems interpretation of the altercentric effect developed by Qureshi et al. (2010): that perspective calculation happens automatically and is informationally encapsulated, while perspective selection is effortful and subject to executive resources. Our adapted DPT was designed to test these predictions.

4.3 Adapted DPT

We address the issues of stimulus type and matching/mismatching responses by conducting a modified version of the DPT in a close conceptual replication of Samson et al. (2010), Experiment 3, and Santiesteban et al. (2014), Experiment 2, using different stimuli and a modified experiment design. Unlike Samson et al. (2010) but following Santiesteban et al. (2014), we used arrows as a directional control for avatars; unlike Santiesteban et al. (2014), we manipulated avatars vs arrows in a between-participants design, rather than within-participants.

The task did not require participants to take the perspective of the avatar, or respond based on how many balls the arrow indicated. Instead, they were required simply to confirm whether the digit on each trial matched the number of red balls in the picture that followed. We predicted that the perspective-taking effect would be found in the Avatar condition, but not in the Arrow condition, based on the hypothesis that previous tasks have found an effect as a result of presenting the two different stimulus types within-subjects or requiring Other trials with inanimate stimuli. We further predicted that, in the Inconsistent condition, No–none trials would be faster than both Yes and No–none trials (because the No–none condition does not require selecting between conflicting perspectives); and that in the Consistent condition, there would be no significant difference between Yes and No–none trials (because neither condition requires perspective selection).

4.3.1 Materials and methods

In order to increase task complexity for a planned series of experiments using multiple avatars simultaneously, we constructed a new set of stimuli using photographs of Lego figures, dubbed "Sally" and "Andrew" for ease of reference, and red beads that, at Lego scale, had the appearance of red balls. Unlike the cartoon avatars used in most DPT implementations, these three-dimensional images had the advantage of unambiguous depth. The barriers were solid black, in order to prevent any ambiguity in whether the Lego characters could see through them. Red balls (the equivalent of the dots/discs in the original DPT) could be placed in view of both of the Lego characters (on the central table); at the foot of each Lego character, visible to that character but not to the other; and behind the characters (and a further black screen), visible to the participant but not to either of the characters (see Figure 4.2). Each scene showed only one avatar and a maximum of four balls, with up to two balls in any given position. Control stimuli consisted of Lego columns with the same colours and proportions as Sally and Andrew, with a black ar-

row on the yellow block at face height, pointing in the same direction as a figure's gaze direction (Figure 4.3).

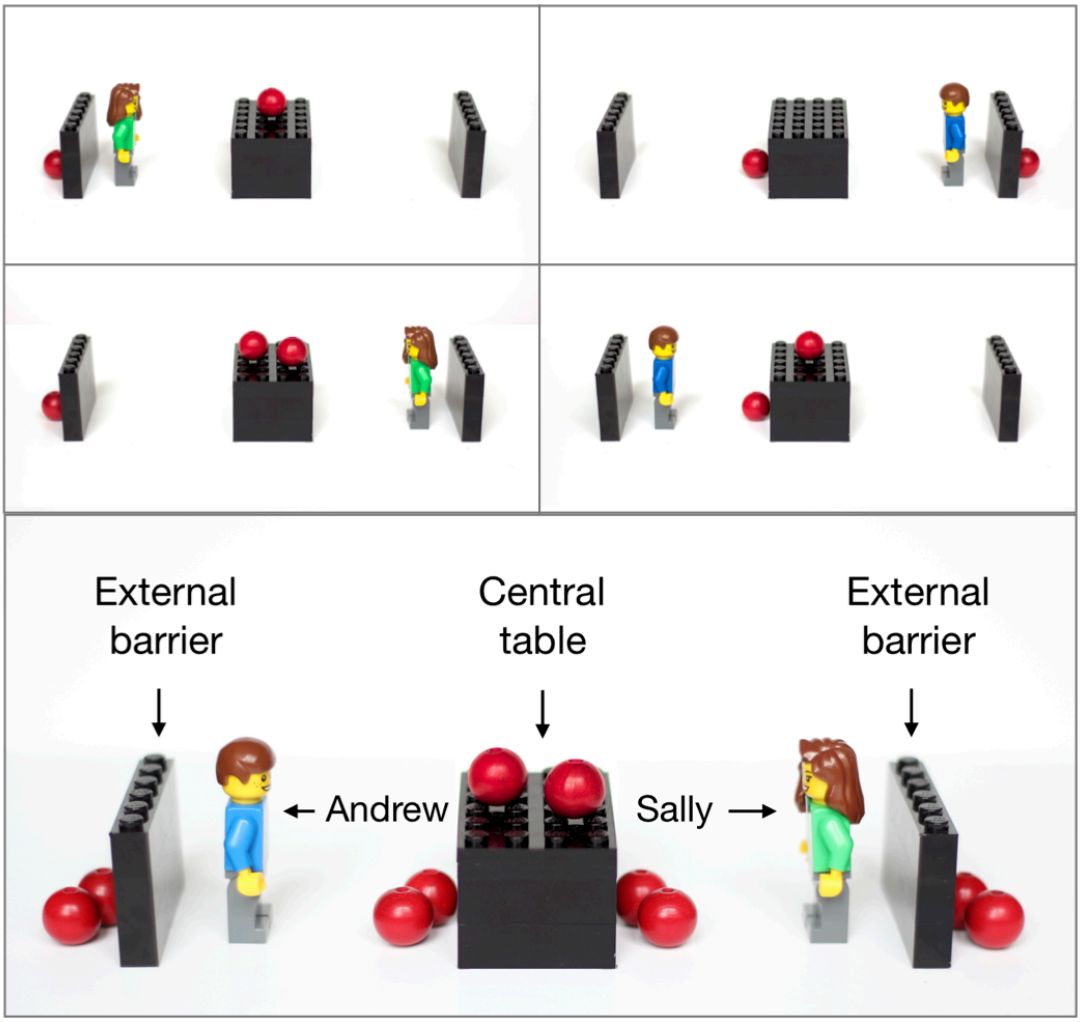


Figure 4.2: Adapted DPT stimuli using Lego figures. The upper four images show example scenes; note that each scene that participants saw featured a single avatar and a maximum of four balls. The lower image shows both potential placement positions for avatars (left or right of the central table) and all possible ball positions (five possible positions, maximum of two balls in any one position).

Participants.

Sixty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 20 minutes. Thirty participants viewed stimuli with the Lego figures, and thirty viewed control stimuli showing columns with arrows on them. One further participant (avatar condition) was excluded from analysis because a post-experiment questionnaire indicated that they

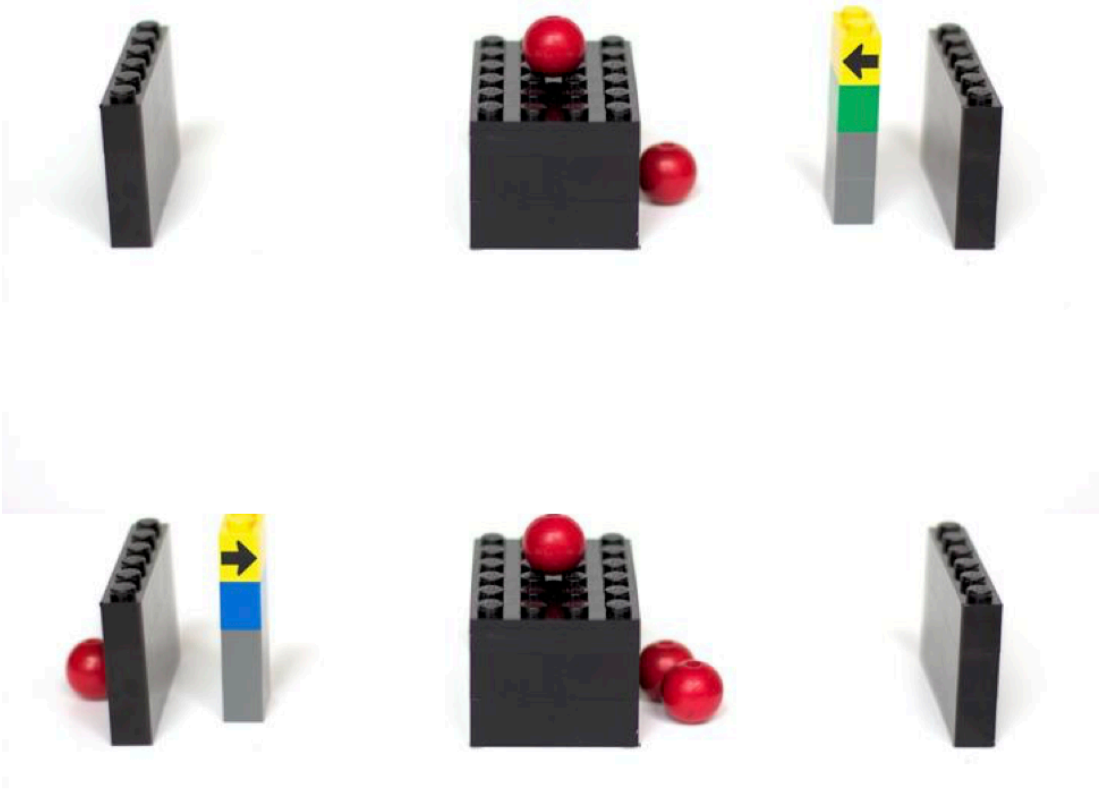


Figure 4.3: Arrow stimuli matching the avatars in both size and colour).

had successfully guessed the purpose of the experiment.

Procedure.

Participants completed a short training sequence explaining the instructions for the task and 12 practice trials with correct/incorrect feedback on responses, followed by 324 trials, divided into four blocks, with a self-paced break between blocks. Their instructions were to judge whether a digit presented before each trial matched the number of balls in the picture that followed. No mention was made of the arrow or avatar stimuli, or of perspective in any sense.

On each trial, participants saw a fixation cross for 750 ms, followed by a digit between 0 and 4 for 750 ms, followed by a Lego scene, with the the words "Yes" and "No" in the bottom corners of the screen (Yes-side was counterbalanced across participants but remained consistent across trials for a given participant). Responses were given using a two-button button box, pressing the Yes-side button when the number of balls matched the digit, and the No-side button when the number of balls did not match the digit. Scenes timed out within 2,000 ms if no response was given, and moved on to the following trial.

The within-subjects variables of interest were the consistency between the audience perspective and the perspective of the stimuli, and the match between the digit shown and the on-screen perspective. Half of all trials were consistent in perspective: that is, the figure/arrow could "see" (i.e. had unobstructed line of sight to) the same number of balls that the participant could see. The other half were inconsistent, with balls hidden from the figure or arrow by either the central, table-like barrier or the external wall-like barriers, introducing an inconsistency between the participant's perspective and that of the avatar/arrow.

One third matched the digit shown, necessitating a "Yes" response from the participant. For example, if the probe digit shown to the participant was "3", and the scene had three balls, the trials was coded as a Yes trial. Two-thirds of trials featured a mismatch between the probe digit and the number of balls in the scene – these are No trials. Within consistent trials, all No trials were, by necessity, No–none. Within inconsistent trials, half of all No trials were No–other, and half were No–none (Figure 4.4).

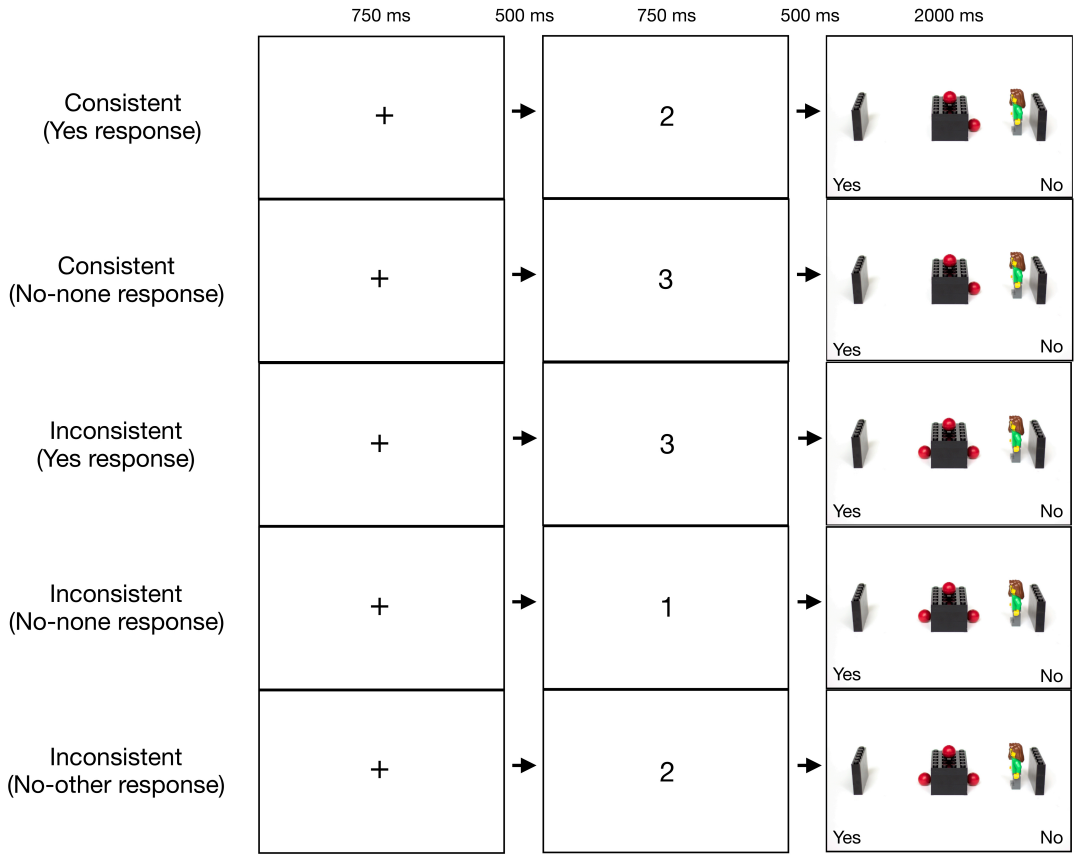


Figure 4.4: An illustration of the trial procedure (fixation cross, digit, scene) in both consistency conditions, with examples of Yes and No–none options in the consistent scenes, and Yes, No–none and No–other options in the inconsistent scenes.

Additionally, a range of other constraints were followed. Each possible number of balls (1, 2, 3 or 4) was presented 72 times. Each avatar, consistency, and yes vs no condition was balanced across the number of balls; that is, there were six No–other inconsistent trials with Sally and one ball present, and six No–none trials in an otherwise identical combination of conditions. In addition, there were 36 filler trials with zero balls. In some cases, this required a particular image to be shown as many as six times; in others, there were multiple image options for a particular combination, in which case six options were randomly selected when the trial list for that participant was generated.

Post-experiment questionnaires were used to assess whether participants' intuitions about the figures' lines of sight matched those of the experimenters. Pictures showing a variety of scenes with balls in different positions were displayed, and participants were asked to note how many balls the Lego figure could see (regardless of whether they had just completed the avatar or arrow condition of the experiment). All responses to these questionnaires indicated that participants did not expect the Lego figures to be able to see balls hidden by either the central or external barriers, but did expect them to see balls either on the table or at their feet.

The experiment was implemented using PsychoPy (Peirce 2010).

4.3.2 Results: planned analysis

We removed training trials, filler trials (those with zero balls), and timed-out trials (0.69%, $n = 119$). No trimming was conducted on higher reaction times, given the imposed cut-off of 2,000 ms on all trials. As per Whelan (2008), trials in which the response RT was lower than 100 ms were also removed, on the assumption that these trials could not be genuine responses to the stimuli (0.02%, $n = 4$). Incorrect responses (3.09%, $n = 531$) were removed and accuracy was analysed separately from RT. Visual inspection of the reaction time data revealed an obvious deviation from the normal distribution, necessitating a log transform of the data (Baayen and Milin 2010). Although log-transformed RTs were used for the analyses, we also report slope estimates in milliseconds, and plot raw RT means, for the sake of clarity.

We used `lme4` (Bates et al. 2015) and `afex` (Singmann et al. 2017) to perform a series of mixed effects regression analyses on the log-transformed reaction times (logRT). We used the standard $p < .05$ criterion for determining where effects were significant, with p-values obtained using model comparison (likelihood ratio tests) using the `mixed()` function in the `afex` package (Singmann et al. 2017) in R (R Core Team 2015).

Table 4.1: RT across consistency for arrows and avatars.

Slope	β	SE	χ^2	df	p
Consistency	0.052	0.011	19.66	1	$< .001^{***}$
Stimulus	0.059	0.048	1.50	1	.22
Stimulus x Consistency	0.01	0.008	1.39	1	.24

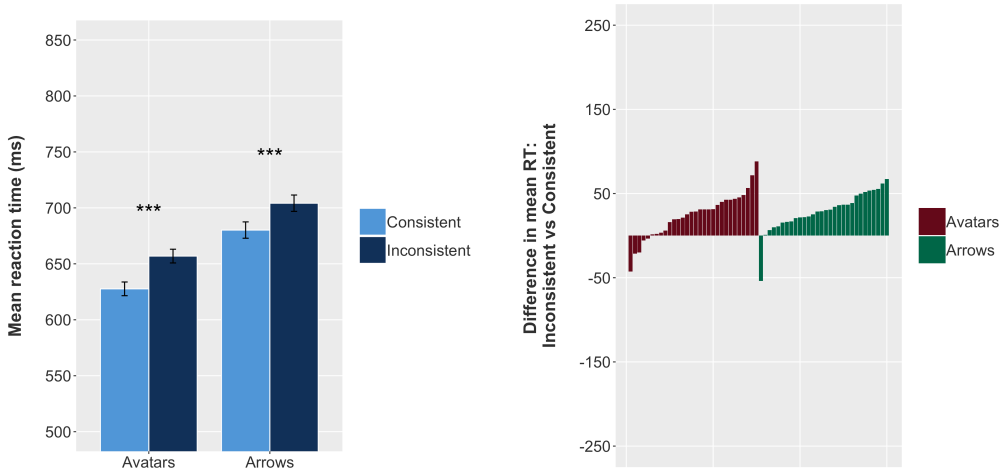
Stimulus

We first conducted a mixed effects linear regression analysis of the relationship between logRT, Consistency and Stimulus. Consistency and Stimulus were deviation-coded (with consistent trials and trials involving avatars rather than arrows being coded positively). As fixed effects, we entered Consistency and Stimulus (with interaction term) into the model. As random effects, we included random intercepts for participants and images, as well as by-participant random slopes for the effect of Consistency.[2]

The model showed an effect of Consistency, suggesting that consistent trials were approximately 33.25 ms faster on average than inconsistent trials (Table 4.1). The difference between estimated RT intercepts for the two Stimulus slopes was of similar magnitude (37.06 ms), but was not significant; this may be explained by inadequate sample size for a between-subjects comparison. There was no Stimulus $\times$ Consistency interaction (see Figure 4.5). This result replicates the core finding of a consistency effect found in previous research, suggesting such an effect for both arrows and avatars. However, an exploratory analysis suggests that this result is purely the result of spatial artefacts in the stimuli, and that there was no perspective-taking or directional effect actually at play. This exploratory analysis, and the theoretical motivation that drove it, are discussed in Section 4.4.

A binomial logistic regression was conducted on error rate, although performance on the task was nearly at ceiling (96.91% accuracy). As fixed effects, we entered Consistency and Stimulus (with interaction term) into the model. As random effects, we included random intercepts for participants and images, as well as by-participant random slopes for the effect of Consistency.[3] This model failed to converge, presumably due to a lack of data, with an average of 8.85 incorrect trials per participant. The by-participant random slopes were dropped (retaining by-participant random intercepts), leading to a less conservative model that did converge,[4] showing no effect on accuracy of either consistency or stimulus, and no interaction (Table 4.2). This replicates

[2]Model syntax: `logRT ~ Stimulus*Consistency + (1+Consistency|Participant) + (1|Image)`
[3]Model syntax: `Accuracy ~ Stimulus*Consistency + (1+Consistency|Participant) + (1|Image)`
[4]Model syntax: `Accuracy ~ Stimulus*Consistency + (1|Participant) + (1|Image)`



(a) (b)

Figure 4.5: Consistency effects for arrows and avatars. (a) Mean RT for consistent and inconsistent conditions, for both arrows and avatars, with y-axis limited to allow direct visual comparison with previous DPT studies; error bars indicate 95% CIs on the mean of the by-participant means. (b) Each individual participant's difference between mean inconsistent RT and mean consistent RT; lines extending above 0 on the y-axis indicate that the participant was slower in inconsistent than in consistent trials. Mean reaction time is higher (i.e. participants respond more slowly) for inconsistent trials for both arrow and avatar stimuli (a); a substantial majority of participants in both conditions show this effect (b).

Table 4.2: Accuracy across consistency for arrows and avatars.

Slope	β	SE	χ^2	df	p
Consistency	0.042	0.044	0.84	1	.36
Stimulus	−0.022	0.122	0.00	1	.96
Stimulus x Consistency	−0.005	0.001	1.39	1	.93

the null results on analysis of errors in the experiments that this study aimed to replicate – Experiment 3 of Samson et al. (2010) and Experiment 2 of Santiesteban et al. (2014). Errors were therefore removed from data for future analyses, and not analysed.

Match

Given that there was no effect of Stimulus in the previous analysis and that avatars are the main stimulus of interest, the analysis of Match was conducted only on the group of participants who saw avatar stimuli, in order to avoid multiple comparisons without any prior hypothesis about arrow-avatar differences in this variable.

Because Match was not balanced across consistent and inconsistent trials (due to the fact that No–other is an impossibility in consistent trials), separate analyses were conducted on con-

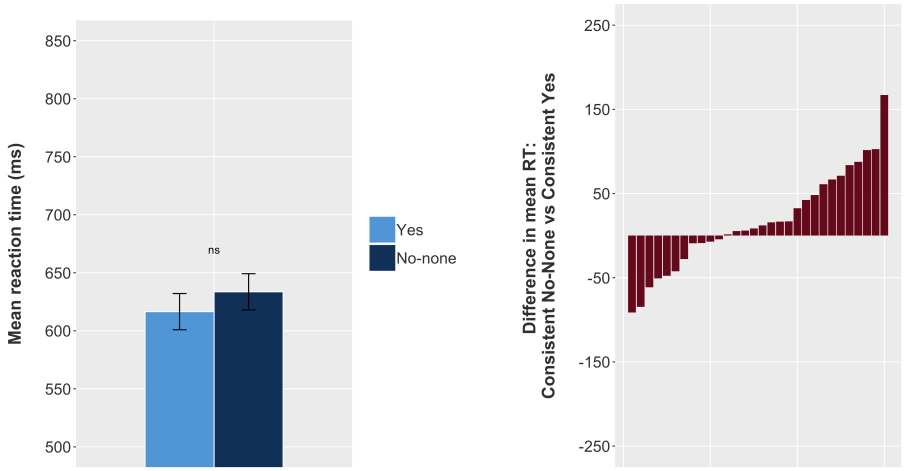
Table 4.3: Effects of Match in Consistent trials.

Slope	β	SE	χ^2	df	p
Match	0.010	0.007	1.67	1	.20

Table 4.4: Effects of Match in Inconsistent trials.

Slope	β	SE	χ^2	df	p
Yes vs No–other	0.009	0.009	4.06	2	.13
Yes vs No–none	0.006	0.007	4.06	2	.13

sistent and inconsistent trials. On consistent trials, Match was sum-coded and entered as a fixed effect into a model with random intercepts for participants and images, as well as by-participant random slopes for the effect of Match.[5] The model showed no effect of Match (Table 4.3), suggesting that, on consistent trials, there was no difference in response times between Yes and No–none trials (Figure 4.6).



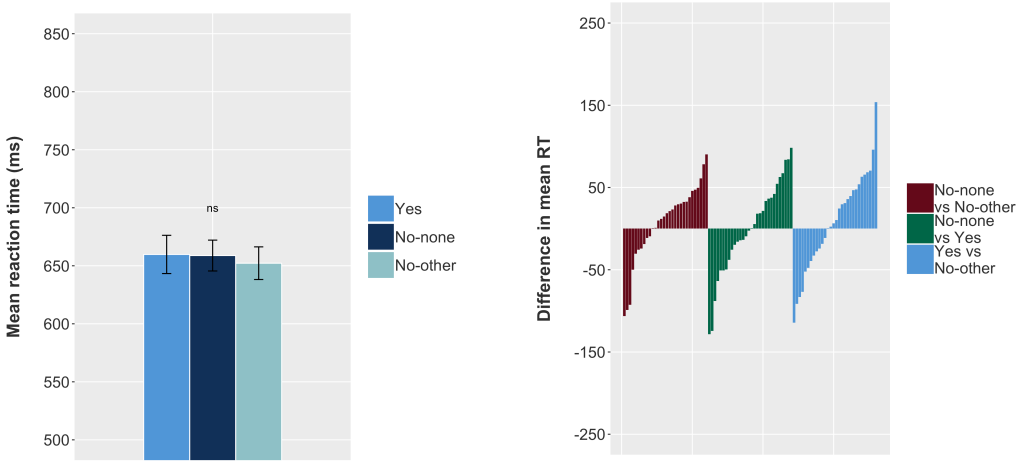
(a) (b)

Figure 4.6: Results of Match analysis, Consistent trials only. (a) Mean RT for Yes and No Consistent trials; error bars indicate 95% CIs on the mean of the by-participant means. (b) Each individual participant's difference between mean No–none RT and Yes RT; lines extending above 0 on the y-axis indicate that the participant was slower in No–none than in Yes trials. A slight majority of participants responded more slowly in No–none trials, but this effect was not significant.

A similar model was run for Inconsistent trials, with two sum-coded slopes for Match, resulting in comparisons between No–none and Yes, and No–none and No–other. There was no effect of Match (Table 4.4, Figure 4.7).

These results do not match the prediction that No–none, entailing no perspective selection,

[5]Model syntax: `logRT ~ Match + (1+Match|Participant) + (1|Image)`



(a) (b)

Figure 4.7: Results of Match analysis, Inconsistent trials only. (a) Mean RT for Yes, No–none and No–other Inconsistent trials; error bars indicate 95% CIs on the mean of the by-participant means. (b) Each individual participant's difference between mean No–none and No–other RT; No–none and Yes RT; and Yes and No–other RT. There was no significant difference in RTs.

should be faster than No–other and Yes Inconsistent trials. They do match the prediction that Yes and No–none should not differ on Consistent trials. Together, these results suggest that there is no meaningful difference in difficulty between the three options.

4.3.3 Discussion

This conceptual replication of the DPT found the headline result of an altercentric effect for both arrows and avatars, with no difference between the magnitude of the effect for the two stimulus types. An additional analysis of the results of Match found no difference between Yes and No in consistent scenes, and no difference between Yes and the two categories of No response in inconsistent scenes, providing evidence against that the perspective-selection account described by Qureshi et al. (2010), and suggesting that with no meaningful difference between these trial types, there is no reason to discard the data from No trials in future implementations of the task.

Although these results seem to support the directional hypothesis, there are alternative interpretations to be considered for the results of not just this DPT adaptation, but all DPT variants using non-social control stimuli. First, an arrow may not be a truly non-social cue. Although arrows have been used as a non-social baseline in gaze direction research (Bayliss and Tipper 2006; Ristic et al. 2002) as well as in the DPT, an arrow is a conventionalised cultural symbol used to direct a viewer's attention, and is in this sense inherently social. It should not therefore be particularly surprising that directing attention is precisely what an arrow does. Further re-

search using directional but truly non-social stimuli as controls may be useful in resolving the question of whether arrows behave more like avatars, or more like non-social directional stimuli. However, even "non-social" stimuli like lamps (Nielsen et al. 2015) and cameras (Wilson et al. 2017) do also carry social meaning: lighting is conventionally used to draw attention (as in the case of spotlights), and cameras are positioned intentionally precisely to capture a particular perspective. The effects of social cueing and directionality may therefore be impossible to tease apart using this method.

Another concern is that demonstrating the altercentric effect with inanimate or non-social stimuli may invite the opposite conclusion: rather than suggesting that the effect does not demonstrate perspective-taking because the effect appears in cartoon arrows and cameras, it could be argued to provide evidence the perspective-taking abilities deployed to understand the perspectives of other humans may be co-opted to imagine the perspective of non-human agents and objects, including cartoon avatars, pets, fictional characters, animate triangles, and so on.

It could also be the case that arrows and avatars produce a similar effect on reaction time, but driven by different processes: visual perspective-taking in the case of avatars, and directional orienting in the case of arrows (Cole et al. 2016). Indeed, Marotta et al. (2012) find that, while eye gaze cues participants to a specific location, an arrow provides a more general cue, suggesting that different processes may be involved in following the directional cue of an arrow and an avatar. The onset of the gaze-cueing effect also appears earlier developmentally than the orienting created by other kinds of stimuli (Farroni et al. 2000; Jakobsen et al. 2013).

One way to discriminate between the effect created by the arrows and that created by the avatars is to establish whether the barriers in front of the stimuli make a difference to what is treated as "consistent" by participants. That is, as in Marotta et al. (2012), if everything in front of the stimulus, regardless of barriers, is treated as "consistent" in terms of reaction time, this would suggest that stimulus is providing a general orienting effect towards everything that it faces. Similarly, if only those balls directly "visible" to the stimulus are treated as "consistent", this would indicate that participants are calculating the perspective of the stimulus. If there is a difference between arrows and avatars, this suggests different underlying processes for the different stimulus. Our experimental design allowed us to conduct an exploratory analysis of this hypothesis, and we predict, in line with Marotta et al. (2012), that the altercentric effect found with arrows will be best explained by directional orienting, while the effect with avatars will be best explained by perspective-taking.

4.4 Directional cueing vs perspective-taking: an exploratory analysis

We conducted an exploratory analysis of the effect of barriers in the data from the adapted DPT, investigating whether the ball locations that were processed as "consistent" differed between arrows and avatars. In order to do this, we re-coded the data, creating a new Consistency variable that accounted for two different definitions of consistency. Scenes in which all the balls appear in the direction faced by the avatar/arrow, regardless of barriers, are "directionally consistent", while those where balls appear behind the avatar/arrow are "directionally inconsistent". Scenes in which all balls are placed on the same side of the central table as the arrow/avatar (i.e. can be "seen") are "line-of-sight consistent". When some balls are hidden from the arrow/avatar but can be seen by the audience, these are "line-of-sight inconsistent". Scenes may therefore be consistent by both definitions (*Avatar sees*), inconsistent by both definitions (*Behind avatar*), or line-of-sight inconsistent but directionally consistent (*Avatar faces*). Line-of-sight consistent, directionally inconsistent scenes are not logically possible. See Figure 4.8 for an illustration of these conditions.

Although previous research (Samson et al. 2010) controlled for the spatial layout of the room, confirming that the presence of the avatar and not merely the distance between the red dots was driving altercentric interference, it is possible that the greater complexity of our Lego scenes could introduce spatial artifacts. Specifically, scenes either have balls clustered entirely around the central table, or include balls on the periphery of the scene, outside the external walls. *Avatar sees* scenes are necessarily central; *Behind avatar* scenes are necessarily peripheral. *Avatar faces* scenes, though, are mixed: some have balls only around the central table, while some include peripheral balls. *Avatar faces* scenes are therefore categorised further into *Avatar faces (central)*, allowing a comparison with *Avatar sees* scenes that controls for the spatial distribution of balls from the centre of the scene; and *Avatar faces (peripheral)*, allowing a spatial distribution-controlled comparison with *Behind avatar* scenes.

It should be noted that this recoding resulted in unbalanced numbers of trials for each new consistency condition, since the experiment had not been designed with this analysis in mind. The definition of consistent trials (*Avatar sees*) did not change, but inconsistent trials were now divided between *Avatar faces (central)*, *Avatar faces (peripheral)*, and *Behind avatar* (see Table 4.5 for the total number of trials in each condition, as well as by-participant means and ranges).

This coding allows an exploratory test of three new predictions:

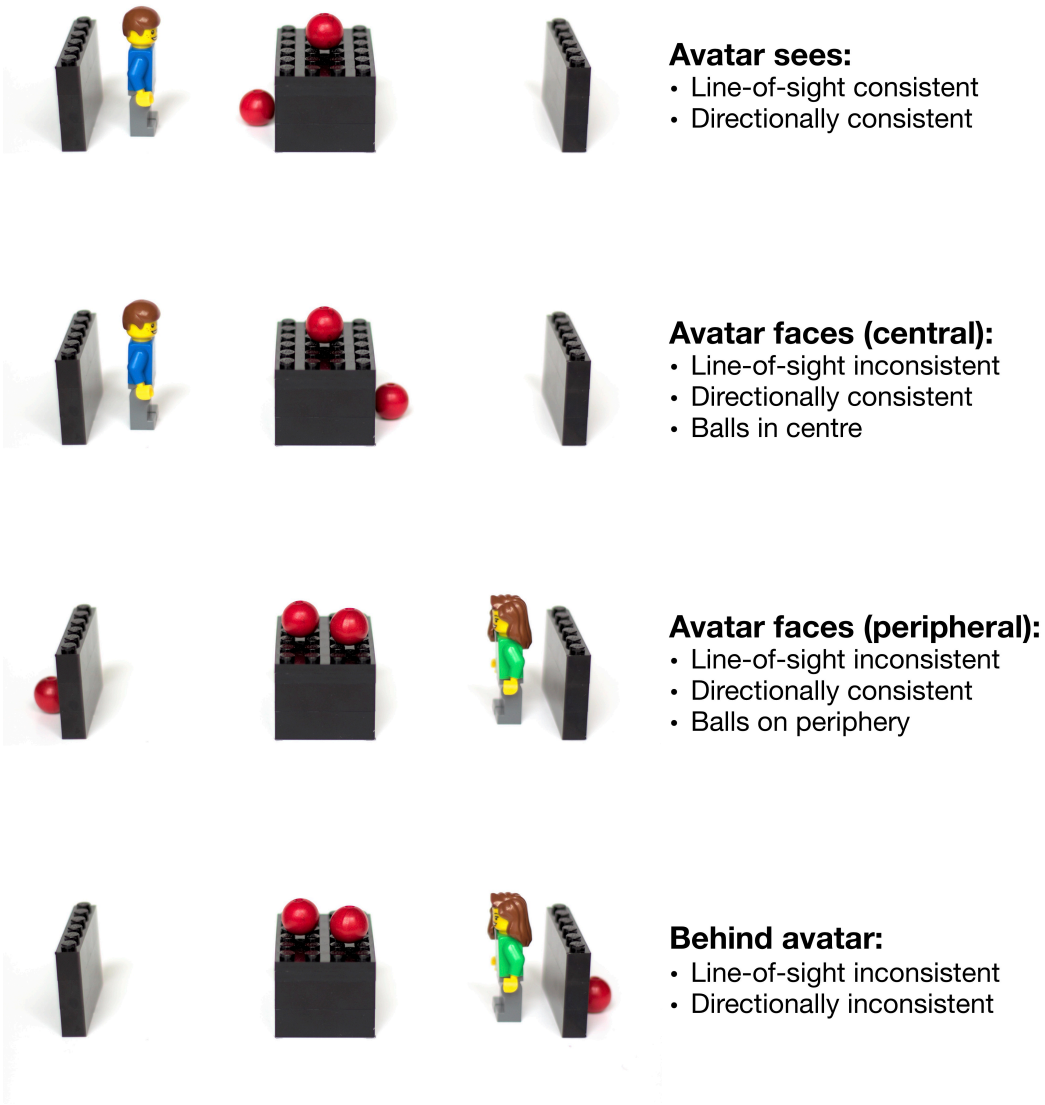


Figure 4.8: Example scenes from the four consistency conditions capturing the differences between avatar and participant perspectives, as well as spatial distribution.

1. As suggested by previous research, *Avatar sees* scenes should be faster than *Behind avatar* scenes, for both arrows and avatars. An effect here would suggest one of three possible causes: it may be due to perspective-taking, directional orienting, or simply a spatial effect, with centrally-clustered scenes processed more quickly than scenes with balls on the periphery. Further comparisons may distinguish between these possibilities.
2. If the effect is driven by perspective-taking, *Avatar sees* scenes should have faster responses than *Avatar faces (central)* scenes in the avatar, since all balls are in the same direction, but only in *Avatar sees* scenes are they truly visible from the avatar's or arrow's perspective. If RTs are equivalent for *Avatar sees* and *Avatar faces (central)* scenes, this implies a direc-

Table 4.5: Trial balancing of redefined Consistency variable in exploratory analysis.

Consistency	Total trials	By-participant mean	Minimum	Maximum
Avatar sees	8,330	138.83	90	144
Avatar faces (central)	1,812	30.20	15	45
Avatar faces (peripheral)	2,696	44.93	30	67
Behind avatar	2,596	43.27	31	63

tional orienting effect. We expect to find an altercentric effect here for avatars, but not for arrows.

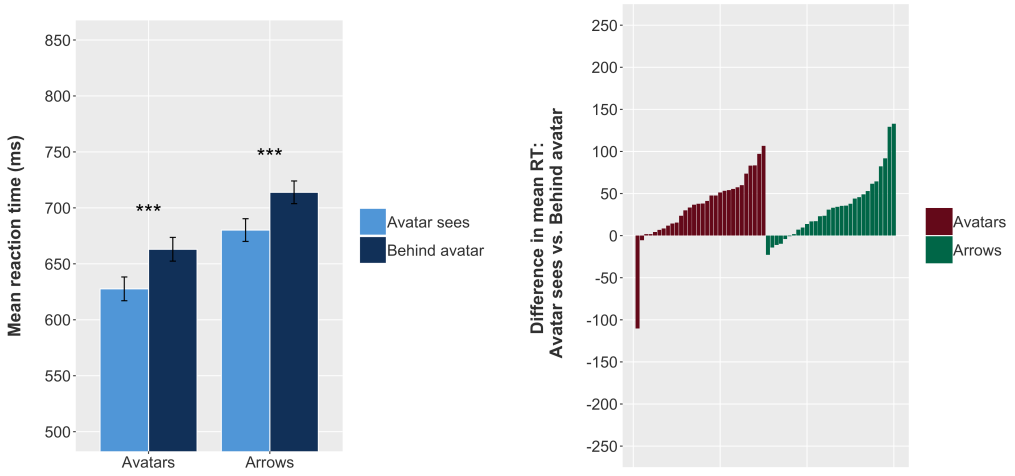
3. If there is a directional orienting effect, there should be faster RTs on *Avatar faces (peripheral)* scenes than for *Behind avatar* scenes. That is, there should be slower responses on scenes involving peripheral balls where some of those balls are behind the avatar, compared to scenes with peripheral balls where some of those balls are in the direction the avatar faces, albeit occluded. On the other hand, if *Behind avatar* and *Avatar faces (peripheral)* scenes have equivalent RTs, this suggests a lack of directional orienting effect, implying that any effect found in comparison 1 is due to spatial artefacts. There is interplay between the results of this comparison and the *Avatar sees* vs *Avatar faces (central)* comparison noted above: if there is an effect in both comparisons, this implies a directional orienting alongside the perspective-taking effect, possibly contributing to part of the delay caused by perspective-taking. If there is a directional effect but not a perspective-taking effect, this implies that any effect found in comparison 1 is due to directional orienting. And if there is no directional orienting effect but there is a perspective-taking effect, it implies that the role of directional orienting in the consistency effect is null or minimal. We expect both arrows and avatars to show a directional orienting effect, but (as discussed above), only avatars to have both a directional orienting and perspective-taking effect.

Prediction 1

We limited the data to *Avatar sees* and *Behind avatar* scenes. Using this subsetted data, we modelled the relationship between Consistency, Stimulus, and logRT. A model with sum-coded Stimulus and Consistency showed a significant effect for Consistency (38.27 ms), no effect of Stimulus, and no interaction between Stimulus and Consistency (Table 4.6, Figure 4.9). The analyses that follows discriminated between three possible explanations for this effect: spatial artefacts, directional orienting, and perspective-taking.

Table 4.6: *Avatar sees* vs *Behind avatar*.

Slope	β	SE	χ^2	df	p
Consistency	0.030	0.004	41.56	1	$< .001^{***}$
Stimulus	0.030	0.024	1.62	1	.20
Stimulus x Consistency	−0.001	0.003	0.20	1	.65



(a) (b)

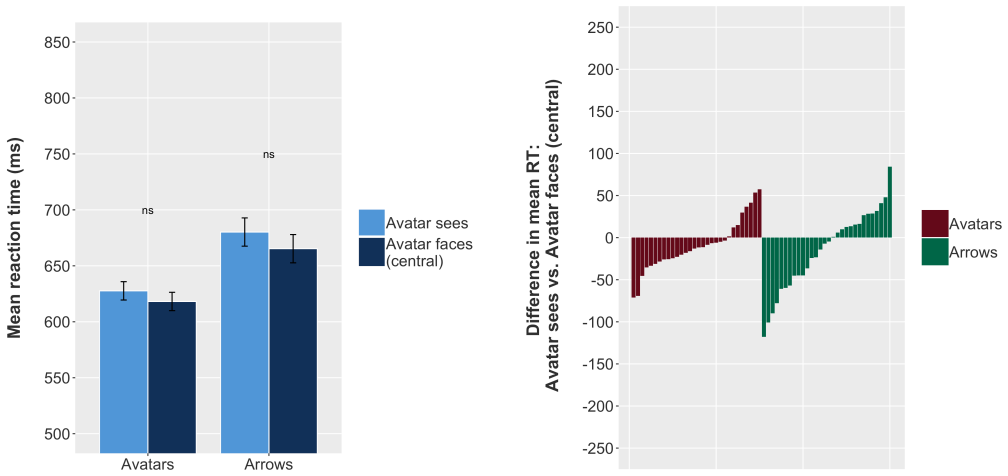
Figure 4.9: Effects of *Behind avatar* vs *Avatar sees* comparison. (a) Mean RT for *Behind avatar* and *Avatar sees* conditions, for arrows and avatars; error bars indicate 95% CIs on the mean of the by-participant means, and significance annotations on the plots reflect the planned comparisons showing the effect of consistency for each condition. (b) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT; lines extending above 0 on the y-axis indicate that the participant was slower in *Behind avatar* than in *Avatar sees* trials (i.e. exhibited an altercentric interference-like effect), while lines extending below 0 indicate that the participant was slower in *Avatar sees* than in *Behind avatar* trials. Mean reaction time is higher (i.e. participants respond more slowly) for *Behind avatar* trials for both stimuli (a); a substantial majority of participants in all three conditions show this effect (b).

Prediction 2

An identical model on a subset of data limited to *Avatar sees* and *Avatar faces (central)* scenes showed no effect of Consistency, no effect of Stimulus, and no interaction between Stimulus and Consistency (Table 4.7). This suggests that, as predicted, arrows do not show a perspective-taking effect, but contrary to prediction, avatars also do not show a perspective-taking effect (Figure 4.10). The effect found in Prediction 1 must therefore be explained by spatial artefact or directional orienting.

Table 4.7: *Avatar sees* vs *Avatar faces (central)*

Slope	β	SE	χ^2	df	p
Consistency	−0.006	0.004	2.17	1	.14
Stimulus	0.031	0.025	1.66	1	.20
Stimulus x Consistency	−0.0002	0.034	0.00	1	.95



(a) (b)

Figure 4.10: Effects of *Avatar faces (central)* vs *Avatar sees* comparison. (a) Mean RT for *Avatar faces (central)* and *Avatar sees* conditions, for arrows and avatars; error bars indicate 95% CIs on the mean of the by-participant means. (b) Each individual participant's difference between mean *Avatar sees* RT and mean *Avatar faces (central)* RT. There is no significant difference for either arrows or avatars (a), and this lack of effect is reflected in individual participant scores (b).

Prediction 3

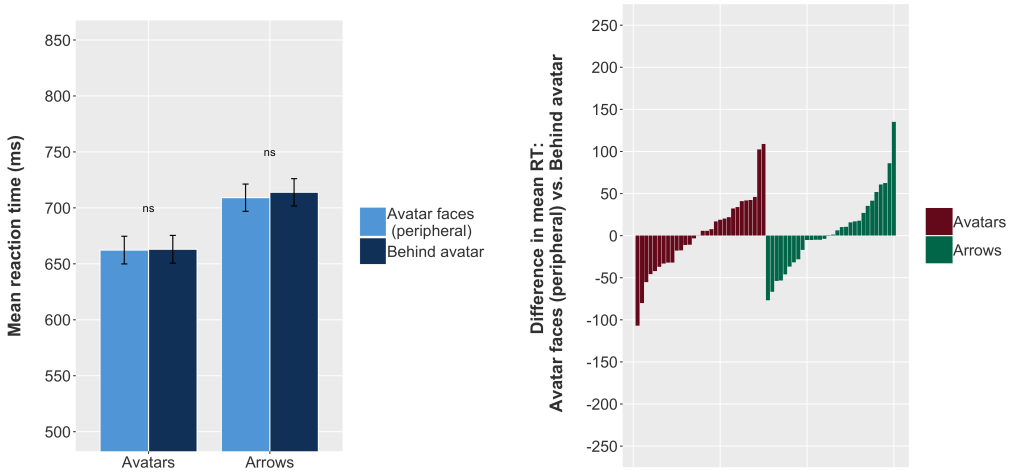
An identical model on a subset of data limited to *Avatar faces (peripheral)* and *Behind avatar* scenes showed no effect of Consistency, no effect of Stimulus, and no interaction between Stimulus and Consistency (Table 4.8). This suggests that neither avatars nor arrows caused directional orienting, and that the consistency effects found in Prediction 1, as well the effects found in the planned analysis, were the result of a spatial artefact (Figure 4.11).

4.4.1 Discussion

These results that suggest that the consistency effect found in the pilot experiment could be explained entirely by the spatial layout of the scene, for both arrows and avatars – that is, neither arrows nor avatars showed either a perspective-taking or directional orienting effect. The lack of altercentric effect after removing this spatial artifact suggests that there was no automatic

Table 4.8: *Avatar faces (peripheral)* vs *Behind avatar*.

Slope	β	SE	χ^2	df	p
Consistency	0.006	0.005	1.39	1	.24
Stimulus	0.027	0.024	1.24	1	.26
Stimulus x Consistency	0.002	0.004	0.21	1	.64



(a) (b)

Figure 4.11: Effects of *Behind avatar* vs *Avatar faces (peripheral)* comparison. (a) Mean RT for *Behind avatar* and *Avatar faces (peripheral)* conditions, for arrows and avatars; error bars indicate 95% CIs on the mean of the by-participant means. (b) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar faces (peripheral)* RT. There is no significant difference for either arrows or avatars (a), and this lack of effect is reflected in individual participant scores (b).

perspective-taking or automatic directional orienting in this task. This conflicts with evidence of both automatic perspective-taking (Samson et al. 2010) and automatic directional orienting (Santiesteban et al. 2014). Possible explanations for this difference in results are discussed in the next section.

4.5 General discussion and conclusions

4.5.1 Automaticity and spontaneity

In their discussion of the initial finding of altercentric interference, Samson et al. (2010, p. 1264) intentionally steer clear of the claim that the results suggest "automatic" perspective-taking:

> *"We have avoided referring to these simple perspective-taking processes as "automatic" because automaticity is difficult to establish without a more exhaustive investigation of*

the range of circumstances in which such phenomena are observed. However, we note that the current data suggest that these processes are relatively resistant to strategic control, with participants computing the avatar's perspective even when they do not need to."

Qureshi et al. (2010) explicitly made the claim of automaticity, at least for perspective calculation (but not for perspective selection), on the grounds that the distraction task engaging executive resources did not remove the altercentric effect, suggesting that these resources were not required to calculate the avatar's perspective. Low use of executive resources is one of three criteria identified as central to automaticity by Bukowski et al. (2015):

1. *Lack of intention.* This is arguably the criterion least open for dispute in the DPT, because the altercentric effect is illustrative of precisely such a lack of intention, in that the avatar's perspective appears to be calculated even when it is not relevant on a given trial.
2. *Low use of resources.* The finding of altercentric interference in a dual-task situation (Qureshi et al. 2010) appears to fulfil this criterion.
3. *Driven reflexively by the mere presence of the stimulus.* If altercentric interference is to be considered automatic by this definition, it should arise even when taking the avatar's perspective is not necessary at all to the entire task – that is, even when the participant is never required to take the avatar's perspective. This appears to be the case in Samson et al. (2010, Experiment 3) and Santiesteban et al. (2014, Experiment 2).

It should be noted that the source of these three criteria identified by Bukowski et al. (2015) argues that automaticity is best conceived of as a matter of degree rather than as a binary, because features themselves may be inherently gradual, and because multiple features combine to contribute to the extent of automaticity of a given process (Moors and De Houwer 2006). There are also alternative accounts of what constitutes automaticity; for instance, Schneider et al. (2017, p. 28) following Bargh (1994) and Shiffrin and Schneider (1977) considers an automatic process to be one that is "unintentional, unconscious, uncontrollable, and attentionally efficient."

Despite this flexibility in what might be considered a truly automatic process, the assumption that automaticity is limited to process that are driven reflexively by the mere presence of the stimulus is widespread in the DPT literature (Cole et al. 2016; Cole et al. 2017; Gardner et al. 2018a; Bukowski et al. 2015; Gardner et al. 2018a), and so I will adopt this as an essential criterion of automaticity.

A range of tasks has explored the question of whether the altercentric effect is purely stimulus-

driven, and can therefore be considered truly automatic. Much of this research draws on the gaze-cueing paradigm (Posner 1980), which has considerable similarities with the DPT. Gaze-cueing studies present participants with an object that can appear on one or the other side of the screen, and measure the participant's reaction time in detecting the object. Between the two sides of the screen where the object appears is a task-irrelevant face, with eyes that look either towards the object to be detected, or away from it. Participants are faster to react on "congruent" trials, in which the face's eye gaze predicts the appearance of the object, than "incongruent" trials, in which the face's eye gaze looks towards the opposite side of the screen.

Evidence from this paradigm suggests that the gaze-cueing effect is not purely reflexive, since it is present only when there is a delay between the appearance of the gaze cue (the face) and the object to be detected. If the two are presented simultaneously, there is no effect; a "Stimulus Onset Asynchrony" (SOA) of more than 50 ms is required to produce the gaze cueing effect (Xu et al. 2012; Frischen et al. 2007). In the DPT, however, the effect has been found when the avatar and discs appear simultaneously. This discrepancy implies that these two otherwise similar tasks have some crucial differences – where effects in the DPT appear to arise immediately, effects in the gaze cueing paradigm appear to require time for participants to process the gaze cue before it becomes effective at directing their attention.

Because the two paradigms have a number of potentially important differences – in the appearance of the stimuli, trial procedure, and experiment design – there are various possibilities that could explain this discrepancy. One possibility is that the brightly coloured avatar is more visually salient than the standard black line drawing of a face used in gaze-cueing task; another is that the potential for a more powerful cueing effect from an entire body, rather than the eye movement on a face (Bukowski et al. 2015).

One of the most crucial differences is that, in many variations of the DPT, the avatar's perspective is either explicitly relevant throughout the task, or mentioned frequently enough that it may become salient. For instance, in Samson et al. (2010, Experiments 1 and 2) and Santiesteban et al. (2014, Experiment 1), participants are required to take the avatar's perspective throughout the experiment. Even when the task is limited to require the participant to respond based only on their perspective throughout, as in Samson et al. (2010, Experiment 3) and Santiesteban et al. (2014, Experiment 2), participants' attention is drawn to the stimulus when, in the task instructions, they are told to ignore it. The likelihood of participants therefore noticing the stimulus, wondering about the reason for its presence if they are to ignore it, and potentially paying it undue attention is then increased by the use of the cue YOU before every trial. Instructing

participants to focus on their own perspective has been found to raise awareness of alternative perspectives by delineating the participant's perspective as distinct from others, making it possible that both the instruction to ignore the avatar's perspective, and repeated pre-trial reminder to answer based on YOU could have unintentionally induced the effect (Bukowski et al. 2015, page 1, 025). This may also induce an "ironic error", which is an error that is made as a result of focusing on not making that error (Wegner et al. 1998).

Bukowski et al. (2015) investigated these possibilities in a series of tasks based on the DPT. The first of these tasks manipulated the SOA of the avatar and dots in the scene, with a simplified design that required participants to judge only whether each scene showed one or two discs, rather than the 0-3 discs and Yes/No responses of the standard DPT. The discs were positioned in front of the avatar ("congruent" scenes) or behind the avatar ("incongruent" scenes) to more closely mimic the gaze cueing paradigm. Participants were never required to take the avatar's perspective, the avatar's perspective was not mentioned, and there was no "YOU" cue before each trial. On trials with SOA of 300 ms, but not those with 0 ms SOA, responses were significantly slower on incongruent than congruent trials – that is, there was an altercentric effect, but only when the avatar appeared before the discs appeared. This result suggests the additional visual saliency of the avatar could not explain the difference between results in the DPT and gaze cueing paradigms; that is, the changes made in this task were sufficient to make the pattern of the DPT's altercentric effect match that of the gaze cueing paradigm, appearing only when the avatar appeared before the dots.

Because this task made a number of adaptations from other DPT variants – including not mentioning the avatar's perspective, and making the task simpler – follow-up tasks investigated which of these changes could account for an altercentric effect having appeared with 0 ms SOA in previous DPT findings. One possible explanation is that the more complex demands of the DPT cause participants to spend longer looking at each scene, resulted in an effect similar to longer SOA by creating extra time for the orientation of the avatar to influence the participant's visual search.

To test this, a second experiment also did not draw attention to the avatar's perspective in any way, but required participants to consider the colour of the dots as well as the number. Again, there was an altercentric effect only on trials with 300 ms SOA, and not those with 0 ms SOA. This suggests that more complex task demands, and more time spent viewing the scene, cannot account for the DPT's altercentric effect at 0 ms SOA.

A third experiment tested the hypothesis that differences in attention, particularly additional

attention drawn to the avatar, could cause the difference. While the gaze-cueing paradigm uses task instructions that draw attention only to the objects in the scene and not the gazer, DPT variants have, as discussed, standardly drawn attention to the avatar through various means.

This third task drew attention to the avatar using different means, in order to focus attention on it without requiring perspective-taking or inducing ironic errors. This was done by changing the trial sequence so that the room and avatar would appear (with no dots) while the trial instructions appeared superimposed on the avatar, with the appearance of the dots controlled separately. The dots appeared either at the same time as the rest of the scene and trial instructions (0 ms SOA) or 300 ms after (300 ms SOA). On this task, there was an altercentric effect at 0 ms SOA, suggesting that drawing attention to the avatar (or potentially any other central stimulus) through any means, including task instructions, may induce computation of the avatar's perspective.

A lack of altercentric effect at 0 ms SOA was replicated by Gardner et al. (2018b) in a task that did not mention perspective in any recruitment materials or instructions; did not require participants to switch perspectives; and did not use the YOU cue before each trial. This task found no altercentric effect for either avatars or an arrow control. A follow-up experiment, closely modelled on a gaze-cueing task, required participants to press a key when they detected the presence of dots appearing in the scene. The dots appeared either in front of or behind the avatar, with arrows used as control stimuli. The altercentric effect appeared for both arrows and avatars at SOA of 600 ms, but not at 300 ms or 0 ms.

These results resolve the apparent inconsistency between the DPT and gaze-cueing literatures by showing that comparable tasks in both paradigms do not result in a reflexive, automatic gaze-cueing effect, but instead require attention to be drawn to the central stimulus in some way to induce an effect. In the DPT, requirements to take the avatar's perspective, or other cues drawing attention to the avatar or its perspective, may be sufficient to create this effect; in the gaze-cueing task or an "uncued" DPT, with no attention drawn to the face or avatar, delayed stimulus onset or superimposed instructions may draw sufficient attention to the gazer to induce an altercentric effect. These results also explain the null effect in the Lego adaptation described in this chapter, which did not require participants to take the perspective of the stimuli and did not mention the stimuli until the post-experiment questionnaire.

Ferguson et al. (2017) used eye-tracking to gain further insight into the altercentric effect. In a standard DPT task using both self- and other trials, they monitored "switch" vs "stick" trials: whether participants switched perspective across two consecutive trials (e.g. "self" followed

by "other") or held the same perspective across both trials (e.g. both "self" trials). They found that the altercentric effect was mediated by task consecutiveness: on "switch" trials, where the previous trial had required the participant to take the avatar's perspective, they found the usual altercentric effect, but on "stick" trials, there was no altercentric effect. The number of fixations on each scene, as determined by eye-tracking, similarly showed an altercentric effect (higher number of fixations on inconsistent trials) on switch, but not stick, trials.

These results all suggest that the altercentric effect is not automatic, according to the definition discussed above; that is, it is not purely reflexive or stimulus-driven, instead arising as a result of attention to the central stimulus, driven by task demands and participants' beliefs about the relevance of the avatar's perspective to the task. This has significance for the "informational encapsulation" claim of the two-systems account, since informationally encapsulated processes are impenetrable to "high-level expectations and beliefs" (Fodor 1983, p. 66). Although informationally encapsulated processes may have access to top-down information flow within a given system (for instance, identification of phonemes making use of lexical knowledge), it is this isolation from the entirety of a person's background beliefs and desires that enables informationally encapsulated processes to perform rapidly (Coltheart 1999). Expectations and beliefs about the relevance of the avatar's perspective to the task, and the resulting desire to take that perspective – which is a clear example of background beliefs and desires – should therefore not affect the calculation of the avatar's perspective if that perspective calculation is part of an informationally encapsulated process; the avatar's perspective should be calculated regardless of participants' beliefs about the task. (See Westra 2017b for a discussion of the role of automaticity in the informational encapsulation claim of the two-systems account.)

Following Westra (2017b), the altercentric effect seems to be better described as a *spontaneous*, rather than *automatic* process (see also Carruthers 2017 and Westra 2017b). That is, where automatic processes are reflexive, cannot be inhibited, and do not depend on attentional processes, goals, or motivation, spontaneous processes may be unconscious, rapid and involuntary, but the deployment of such processes is determined by attention and intention. As an illustration, consider seeing in colour compared to seeing in focus: while seeing in colour is automatic, seeing in focus is controlled by attention, and is therefore spontaneous by this definition.

While these findings challenge a two-systems account of the altercentric effect, the submentalising account could still explain all the findings described so far, leaving open the question of whether the apparently *spontaneous* perspective-taking found in tasks that draw sufficient attention to the avatar is a result of perspective-taking, or of directional orienting. Results from

tasks attempting to answer this question are inconsistent. The following section reviews these results, explaining the possible role of task instructions and demands in explaining the inconsistencies, and identifying opportunities for future research.

4.5.2 Occlusion tasks: the role of task instructions and demands

Although the submentalising explanation of the DPT was first investigated by comparing arrows with avatars, the use of non-humanoid stimuli such as arrows and cameras is not an ideal test of mentalising vs submentalising accounts. This is because it investigates two separate questions simultaneously: first, is the altercentric effect caused by mentally representing the perspective of the avatars, or by preferentially attending to the dots/discs on the side of the screen faced by the avatars? And secondly, is this effect limited to "social" stimuli, or can it be induced by any directional stimuli? Answering the second question is particularly difficult. As discussed above, it may be the case that different underlying processes could create similar results for avatars and arrows, as discussed above – that is, perspective-taking in avatars and directional orienting for arrows (Marotta et al. 2012).

The first question, on the other hand, is more easily answered by obscuring some of the discs from the avatar's vision, creating a context in which the avatar continues to face certain dots (able to induce directional orienting) but is unable to see these dots (unable to induce perspective-taking). The exploratory analysis described above, treating the central table in the Lego as a barrier that allows the Lego character to face but not see certain dots, is one possible way to achieve this. Although this task found a null result, a range of different "occlusion" tasks have used variants of the DPT in which certain dots are placed in front of the avatar but are not visible to the avatar, and found an altercentric effect. However, some of these results support the perspective-taking account of the altercentric effect, and others support the directional orienting account.

These two accounts make separate predictions for occlusion tasks:

1. The perspective-taking account predicts that there will be a difference in RT between scenes in which the avatar faces, but does not see, all the dots (*Avatar faces*) and those in which the avatar faces and sees all the dots (*Avatar sees*). It predicts that there will *not* be a difference in RT between scenes in which dots are hidden behind the avatar (*Behind avatar*) and those in which dots in front of the avatar are hidden by occlusion (*Avatar faces*), because both of these scenes would be inconsistent. These findings would suggest that the altercentric effect is driven by what the avatar sees, rather than what it faces.

2. The directional orienting account predicts that there will not be a difference in RT between scenes in which the avatar faces, but does not see, all the dots (*Avatar faces*) and those in which the avatar faces and sees all the dots (*Avatar sees*). It predicts that there *will* be a difference in RT between scenes in which dots are hidden behind the avatar (*Behind avatar*) and those in which dots in front of the avatar are hidden by occlusion (*Avatar faces*), because scenes in which the avatar faces the dots will be processed faster. These findings would suggest that the altercentric effect is driven by what the avatar faces, rather than what it sees.

Cole et al. (2016) use a barrier in front of the avatar, with a window in the barrier that may be open or closed (see Figure 4.12). When the window is open, the avatar can see the dots in front of it ("seeing" trials); when it is closed, the dots are blocked from the avatar's perspective ("non-seeing" trials). On consistent trials, the avatar would face the same number of dots that the participant could see, but would only be able to "see" these dots through the open window on "seeing" trials (i.e. *Avatar sees* and *Avatar faces*). On inconsistent trials, dots appeared behind the avatar (*Behind avatar*). This task found a difference in RT between *Behind avatar* and *Avatar sees*, which is consistent with both the perspective-taking and directional orienting accounts; and found a difference in RT between *Behind avatar* and *Avatar sees* / *Avatar faces*, suggesting that the altercentric effect was driven by directional orienting, rather than perspective-taking.

Baker et al. (2016) found directly contradictory results in a set of three experiments using barriers. In Experiment 1, avatars placed on the far end of a room faced into the room, which had up to three orbs floating in front of them. A barrier wall could be placed in between these orbs, preventing the avatar from seeing any of the orbs behind the wall. Participants could see all the orbs, and could judge that the avatar was not able to see the orb(s) behind the barrier. On consistent trials, all orbs were on the same side of the barrier as the avatar (*Avatar sees*); on inconsistent trials, some orbs were hidden behind the barrier (*Avatar faces*). If the altercentric effect were driven by directional orienting, there should be no effect in this experiment, since the avatar always faced all of the orbs. However, there was an effect, suggesting that the delay was a result of perspective-taking rather than directional orienting.

In order to confirm that the effect was not due to the spatial layout of the room, Baker et al. (2016) modified the experiment so that all orbs were stacked vertically, with wall segments blocking some of the orbs on some trials. Again, they found an altercentric effect. Finally, a third experiment used a window in the barrier wall to confirm that the avatar's visual access, rather than the placement of the barrier, explained the effect. In a between-subjects design, one

group of participants repeated Experiment 1 (with a solid wall blocking some of the orbs in inconsistent trials), while the second group had a "window" in the barrier wall that made all of the orbs visible on every trial (see Figure 4.12). This experiment is closely comparable with Cole et al. (2016), with the primary difference being the presentation of window and wall conditions as between-subjects rather than within-subjects. In direction contradiction to the results of Cole et al. (2016), however, there was an altercentric effect for the "wall" condition, but not for the "window" condition, suggesting that the avatar's visual perspective, rather than the direction it faced, was driving the altercentric effect.

A variety of differences between the tasks could result in these contradictory findings. One possibility is the temporary nature of the barriers used by Cole et al. (2016): the fact that the agent can sometimes see what is on the other side of the barrier, and at other times cannot, may make it simpler for participants to simulate a consistent perspective for the avatar throughout (that is, seeing the dots) rather than switching between two different possible perspectives.

Another difference is in the visual clarity of the stimuli. While Baker et al. (2016) used stimuli with a clear third dimension that showed unambiguously what the avatars were able to see, the images used by Cole et al. (2016), which are an adaptation of those used in the original DPT, may be open to ambiguous interpretations. It is not clear that the barrier is not transparent; nor, if the barrier is assumed to be opaque, that the "window" is transparent, rather than a screen; and the depth and angle of the barrier placement within the room could be misinterpreted as not appearing between the avatar and the dots. However, measures were taken to ensure that the visual layout of the scene was properly understood: open or closed barriers were shown in different blocks of trials, and participants were explicitly told whether or not the avatar could see the wall that was blocked by the barrier at the beginning of each block. When a separate group of participants was asked to judge what the avatar could see, 100% of their responses were correct.

Perhaps the most important difference is that Baker et al. (2016) required participants to judge both their own perspective and the avatar's perspective – that is, the task explicitly demanded perspective-taking throughout. Cole et al. (2016), on the other hand, never required participants to take the avatar's perspective, instead requiring them only to verify whether the digit presented matched the number of dots in the room. They were told explicitly what the avatars could and could not see at the beginning of each block of trials, implicitly cueing them to the importance of perspective-taking to the task.

Given the evidence discussed in Section 4.5.1 that uncued tasks do not show an altercentric

effect, and that various methods of drawing attention to the avatar (including requiring the avatar's perspective to be used in the task) are successful in inducing this effect, this raises the possibility that results could differ depending on the level of attention, and type of attention, drawn to the avatar. "Explicit" tasks that require participants to take the avatar's perspective on half of all trials may result in this perspective-taking being sustained throughout the task, while "implicit" tasks that prompt a lesser degree of attention, but nonetheless draw some attention to the avatar, may create a lower-level directional orienting effect.

Directional orienting is a necessary input to perspective-taking: in order to take someone's perspective, the direction that they face must be calculated. Perspective-taking then requires further calculations about obstructions in the field of view and the status of the person's vision (eyes open or closed, direction of eye gaze, and so on), suggesting that directional orienting may be a lower-demand precursor to perspective-taking. Since Cole et al. (2016) found evidence consistent with directional orienting on an implicit task, and Baker et al. (2016) found evidence consistent with perspective-taking on an explicit task, this difference in attention drawn to the avatar could explain the conflicting results. A later implicit occlusion task that found results consistent with the directional orienting account (Langton 2018) provides additional support for this hypothesis.

A different series of occlusion tasks has used avatar blinding, rather than barriers, to block the avatar's line of sight. Furlanetto et al. (2016) acquainted participants with two pairs of goggles, one transparent and the other with blocked lenses (either red or orange, with the colour counterbalanced across participants). The on-screen avatars then wore goggles of each colour in different trials, inviting the interpretation that in some trials, the avatars were blinded. This task, like Baker et al. (2016), was explicit, requiring participants to judge both their own perspective and the avatar's perspective. There was an altercentric effect on trials with transparent goggles, but no effect when avatars wore opaque goggles, supporting the perspective-taking account.

Conway et al. (2017) argue that methods using barriers and avatar blinding do not rule out submentalising interpretations that rely on learned associations rules rather than perspective-taking to explain the altercentric effect, because given sufficient experience with barriers in everyday experience, it should be possible for participants to learn a general rule such as "when barriers are placed between an agent and an object, they do not interact with the object." In order to counter this possible interpretation, they create an experiment that separates seeing vs non-seeing trials without using barriers or blinding, instead using a stimulus with which participants were unlikely to have any experience.

In this task, participants were acquainted with two telescopes of different colours (counterbalanced across participants). One of these telescopes allowed viewing of a red dot in a small model room, while the other had an adjusted focal length that rendered the dot invisible. After the induction phase, participants completed a DPT in which the stimuli (both arrows and avatars) were looking through these different-coloured telescopes at the dots in the room. Both stimulus types could appear with either colour of telescope. Participants were instructed to judge whether the number of dots in the room matched the dots they could see, and were cued with the word YOU before each trial, making this an implicit perspective-taking task.

There was an altercentric effect, but it did not vary with the type of telescope, supporting the directional orienting account. In line with Cole et al. (2016) and Baker et al. (2016), these blinding tasks support the pattern that implicit and explicit tasks create different effects, with implicit tasks resulting in an directional orienting effect and explicit tasks resulting in a perspective-taking effect. However, in this case, the telescope task is also potentially complex and unfamiliar enough that an involuntary and rapid computation of other-perspective becomes impossible, causing participants to revert to simpler processes such as directional orienting; the degree of difference between this task and the original DPT make them difficult to compare precisely.

The possible difference in results between explicit and implicit tasks is noted by Conway et al. (2017), who attempted to test whether the implicit or explicit task requirements could be responsible for the results in a replication of the explicit Furlanetto et al. (2016) goggle task. The effect failed to replicate, despite a sample size three times greater than the original task, suggesting that the original finding may have been a false positive. The same participants who completed the replication also completed an implicit goggle task, in which they were required to take only their own perspective. This task found an altercentric effect that was actually greater when avatars were wearing opaque goggles, providing support for the directional hypothesis. However, given the failure of the replicated Furlanetto et al. (2016) task, the fact that these implicit and explicit tasks were within-subjects, and the overall complexity of the goggle task, the interpretation of these results is difficult.

Other tasks using blinding have produced results that are similarly difficult to interpret. Wilson et al. (2017) used a blindfolded or seeing avatar as a between-subjects variable. This was an explicit task, meaning that certain participants only ever saw a blinded avatar, but were nonetheless asked to respond based on what the avatar could see for half of all the trials they completed. It seems likely that the strange instruction to take the avatar's perspective may have overridden the appearance of the blindfolded avatar not being able to see any dots, explaining the appearance

of an altercentric effect for both conditions.

A conceptual replication of Furlanetto et al. (2016) that used an online video to acquaint participants with the transparent and opaque goggles, and an avatar that was not oriented directly towards the dots, found no altercentric effect on reaction times (Marshall et al. 2018). There was an altercentric effect on error rates for avatars with translucent goggles or no goggles, and no altercentric effect for opaque goggles, but given the failure to replicate the effect on reaction times and the multiple comparisons within a three-way repeated-measures ANOVA, this may be a false positive.

The inconsistency of results in occlusion tasks that use blinding methods suggest that these tasks may be difficult to implement with stimuli that are clear enough to be interpreted similarly by all participants, and that are intuitive enough to be incorporated into a task in which responses are usually given in less than a second. For this reason, barrier tasks may provide a clearer picture on whether task presentation plays a role in producing results that are consistent with either the perspective-taking or directional orienting account.

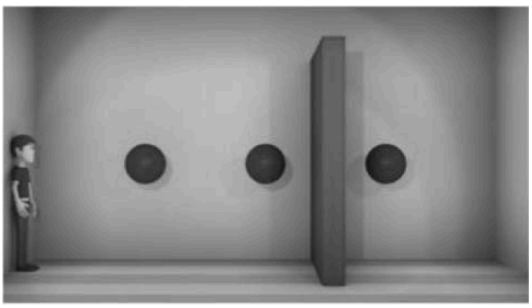
In the following chapter, we present a series of experiments manipulating task presentation (explicit, implicit and uncued) in a between-subjects design, using the Lego stimuli described in this chapter in an occlusion task that is designed to differentiate between balls in the avatar's field of view, compared to those in the direction it faces.

4.6 Chapter summary

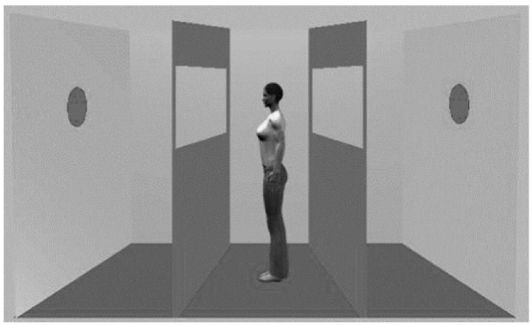
This chapter surveyed a range of DPT variants using "non-social" stimuli such as arrows, cameras and lamps to establish whether the altercentric effect is best explained by mentalising or submentalising. It presented a conceptual replication of (Samson et al. 2010), Experiment 3, and (Santiesteban et al. 2014), Experiment 2, using Lego figurines and size- and colour-controlled arrows. Results from this task showed no difference between arrow and avatar stimuli, and contrary to assumptions across much of the DPT literature, no differences between Yes, No–none, and No–other trials. The results did suggest an altercentric effect, but an exploratory analysis controlling for the spatial layout of the scene suggested that this effect was introduced by a spatial artifact.

A review of DPT variants investigating automaticity suggest that the best explanation for this unexpected null effect may be the fact that this experiment was "uncued", providing participants with no instruction to consider the avatar's perspective as relevant to the task. Previous uncued

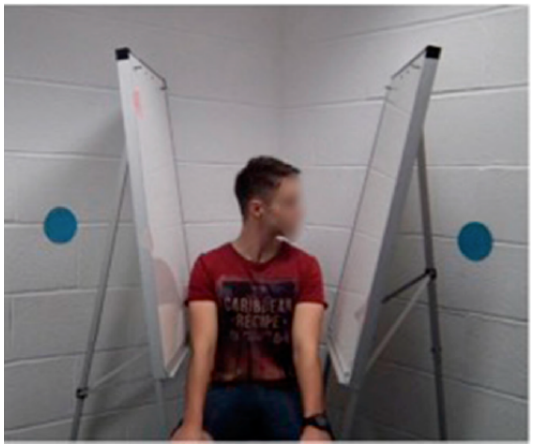
tasks have found no altercentric effect without drawing additional attention to the avatar (Gardner et al. 2018b; Bukowski et al. 2015). A similar methodological difference underlies apparently contradictory results in DPT variants using "occlusion" tasks that screen dots from the avatar's perspective using physical barriers or blinding: the "explicit" tasks that require participants to take the avatar's perspective in the task have shown results consistent with perspective-taking, while "implicit" tasks have shown results consistent with directional orienting. Chapter 5 operationalises this methodological distinction to test mentalising and submentalising interpretations of the task.



(a) Baker et al. (2016), explicit barrier task

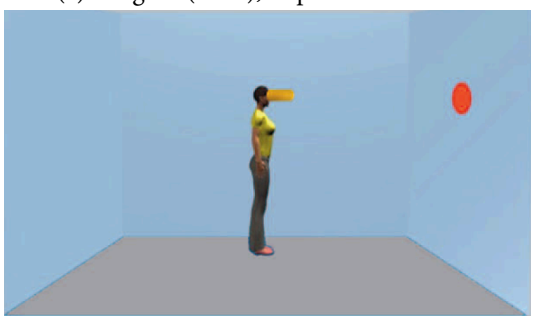


(b) Cole et al. (2016), implicit barrier task



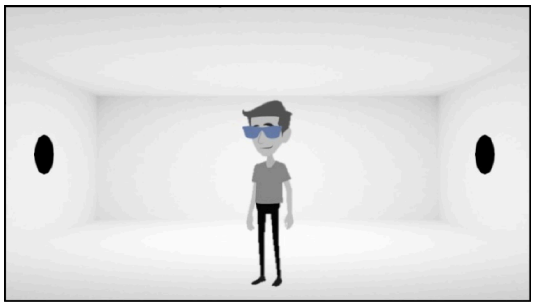
(c) Langton (2018), implicit barrier task

(d) Furlanetto et al. (2013), explicit blinding task (goggles)



(e) Conway et al. (2017), implicit blinding task (telescope)

(f) Wilson et al. (2017), explicit blinding task (blindfold)



(g) Marshall et al. (2018), explicit blinding task (goggles)

Figure 4.12: Examples of stimuli used in occlusion tasks.

Chapter 5

Perspective-taking is spontaneous but not automatic

5.1 O'Grady et al.

Material in this chapter has been reproduced from O'Grady et al., submitted to the *Quarterly Journal of Experimental Psychology*. This paper was co-authored with Thom Scott-Phillips, Suilin Lavelle, and Kenny Smith, and has been reproduced here with the permission of all authors. The research was conceived by all authors, and all authors contributed to the writing and editing of the paper. I conducted the research, and analysed the data with assistance from Kenny Smith. Rachel Kindellan collected the data for Experiment 5. Section 5.2 has been reproduced from the relevant sections in the supplementary material included with O'Grady et al.

Note that the pilot experiment, referred to in this paper as being in the Supplementary Material, is the experiment described in the previous chapter.

Perspective-taking is spontaneous but not automatic

Cathleen O'Grady[1], Thom Scott-Phillips[2, 3], Suilin Lavelle[1] and Kenny Smith[1]

Abstract

Data from a range of different experimental paradigms – in particular (but not only) the dot perspective task – have been interpreted as evidence that humans automatically track the perspective of other individuals. Results from other studies, however, have cast doubt on this interpretation, and some researchers have suggested that phenomena that seem like perspective-taking might instead be the products of simpler behavioural rules. The issue remains unsettled in significant part because different schools of thought, with different theoretical perspectives, implement the experimental tasks in subtly different ways, making direct comparisons difficult. Here, we explore the possibility that subtle differences in experimental method explain otherwise irreconcilable findings in the literature. Across five experiments we show that the classic result in the dot perspective task is not automatic (it is not purely stimulus-driven), but nor is it exclusively the product of simple behavioural rules that do not involve mentalizing. Instead, participants do compute the perspectives of other individuals rapidly, unconsciously and involuntarily, but only when attentional systems prompt them to do so (just as, for instance, the visual system puts external objects into focus only as and when required). This finding prompts us to clearly distinguish spontaneity from automaticity. Spontaneous perspective-taking may be a computationally efficient means of navigating the social world.

Keywords

perspective-taking; dot perspective task; automaticity; spontaneity; directional orienting

Introduction

Everyday interactions with other people seem to require us to keep track of what those around us can see. Actions as simple as asking a friend to hand you an object, passing a football to a team member, or assessing whether an oncoming pedestrian has noticed your bicycle appear to require tracking what another individual can see – that is, visual perspective-taking. Taking the visual perspective of another individual is a form of mindreading, requiring a mental representation of another person's visual field (Apperly 2011). However, it could be the case that behaviours like these are guided by a less complex cognitive process, such as directional orienting, in which an agent is simply aware of what appears in the direction that another individual is facing (Heyes 2014). Currently, much debate on visual perspective-taking centres on the question of whether results in certain visual perspective-taking tasks are better explained by mentalizing or by submentalizing processes such as directional orienting (Conway et al. 2017; Freundlieb et al. 2016, 2018; Gardner et al. 2018b; Langton 2018; Santiesteban et al. 2014; Zhao et al. 2015; Gardner et al. 2018a).

One significant reason why these empirical issues are presently unresolved is methodological inconsistencies in the experimental literature. Despite the fact that much of the literature uses the same basic experimental task (see below) there are nevertheless recurrent variations in experimental design, making truly direct comparisons difficult. As we detail below, one crucial difference is the presence or absence of various prompts cueing participants to consider perspective-taking relevant to the task. This methodological choice is made for a variety of reasons, but especially key are differing assumptions about whether excluding prompts in certain tasks provides a more genuine assessment of spontaneous or automatic perspective-taking (Conway et al. 2017; Gardner et al. 2018a,b; Santiesteban et al. 2014; Bukowski et al. 2015). There is further inconsistency in the use of the terms automatic and spontaneous themselves, which are used interchangeably in some papers, hindering clarity in the debate (Cole et al. 2016, 2017; Langton 2018; Michael et al. 2018).

Here we address these issues. We first present a literature review that summarises the key issues identified above, discussing the utility of making a principled distinction between automatic and spontaneous processes. We then present three new preregistered studies that address the issues directly, using the same experimental task as much of the existing literature (the Dot Perspective Task; see below), and two replications using alternative stimuli. Collectively our results show that one particular variant of the task does indeed demonstrate computation of another individual's perspective; that is, it involves perspective-taking rather than directional orienting. This effect arises rapidly and involuntarily (i.e. it is spontaneous), but it is not found uniformly across

[1]School of Philosophy, Psychology & Language Sciences, University of Edinburgh
[2]Dept. of Cognitive Science, Central European University, Budapest
[3]Dept. of Anthropology, Durham University

Corresponding author:
Cathleen O'Grady, University of Edinburgh, School of Philosophy, Psychology and Language Sciences, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK
Email: C.J.O'Grady@sms.ed.ac.uk

different task designs (i.e. it is not automatic). The effect depends instead on whether the perspective of the avatar (or other stimulus) is made salient in one way or another. We further show that in another variant of the task, effects vary depending on the stimuli used, further corroborating the evidence that responses are not automatic, depending instead on participants' interpretation of the task requirements. Collectively, these results indicate that attentional processes moderate the deployment of perspective-taking. This finding explains apparent inconsistencies in the literature, and suggests that perspective-taking and directional orienting may both play a role in responses, depending on task context.

The Dot Perspective Task

The Dot Perspective Task (DPT) requires participants to enumerate the number of dots that appear in a scene containing an avatar that sometimes has a different perspective from the participant's (see Figure 1 for detailed description). The classic result is that participants are slower to respond based on their own perspective when the avatar's perspective differs from their own (Cole et al. 2016; Conway et al. 2017; Furlanetto et al. 2016; Nielsen et al. 2015; Qureshi et al. 2010; Samson et al. 2010; Santiesteban et al. 2014; Surtees and Apperly 2012). This result is sufficiently well-established that in recent years the DPT has begun to be used to establish the presence or absence of perspective-taking abilities in a range of different contexts, including research on psychopathy and gender differences (Drayton et al. 2018; Yue et al. 2017).

However, the interpretation of results from the DPT is disputed. On the one hand, data from the DPT are often cited as evidence that participants "automatically" (Drayton et al. 2018; Furlanetto et al. 2016; Michael et al. 2018) or "spontaneously" (Gardner et al. 2018b; Samson et al. 2010; Surtees et al. 2016; Cole et al. 2016, 2017) compute the perspective of the avatar. This is because of the robust finding of altercentric interference: the conflicting perspective of the avatar slows down computation of what the participant herself sees. This occurs even on trials when the avatar's perspective is strictly irrelevant to participants' task of responding to the number of dots they (the participant) can see. Since computing the avatar's perspective on these trials runs counter to the task instructions (both the instruction to take the perspective indicated on each trial, and the instruction to respond as rapidly as possible), and since the avatar's perspective is not relevant to calculating the correct answer, the altercentric effect suggests that representation of the avatar's perspective occurs involuntarily on these trials.

On the other hand, some variants of the DPT produce results that motivate an alternate explanation, namely that the altercentric interference effect is caused not by participants taking the perspective of the avatar and being slowed accordingly, but rather by the avatar serving as a directional cue directing participants' attention to certain dots (Cole et al. 2016, 2017; Langton 2018; Santiesteban et al. 2014). That is, altercentric interference may be explained not by participants forming a representation of the avatar's line of sight, but rather by preferentially attending to the dots that the avatar "points" toward.

The next section discusses the various versions of the DPT that have been used to investigate these issues, and the corresponding differences in task design that make the results from various studies difficult to reconcile.

Perspective-taking or directional orienting: Differences in task design

An early modification to the DPT investigated whether altercentric interference would be found for stimuli that had a direction, but no agency of their own* (Santiesteban et al. 2014). This study found altercentric interference not only for avatars, but for arrows too, which was interpreted as evidence that avatars (and arrows) serve as a type of directional stimulus, prompting directional orienting rather than visual perspective-taking itself. There might, however, be different processes involved in each case: visual perspective-taking in the case of the avatar, and directional orienting in the case of the arrows (Cole et al. 2016). Indeed gaze-cueing research suggests that, while eye gaze

*It is of course worth noting that the avatars in the DPT do not have agency themselves, since they are cartoon-like representations of people rather than actual people. But avatars do at least aim to imitate things that do have agency, unlike e.g. arrows.

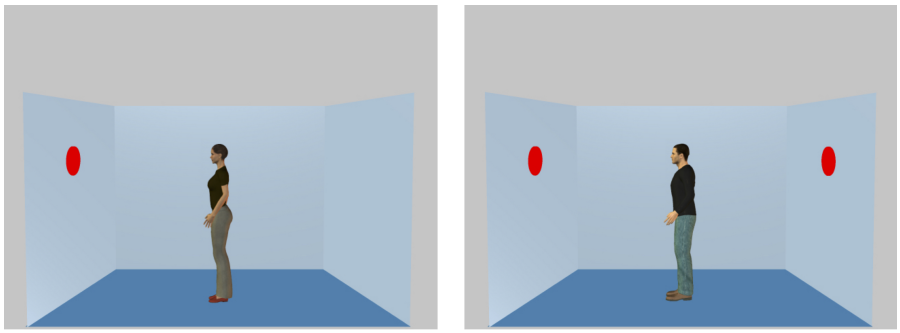


Figure 1. Stimuli from the original DPT (Samson et al. 2010). The task requires participants to view a scene that includes a human avatar and an array of dots. On every trial, participants are told whether to take the avatar's perspective (with the prompt HE or SHE) or their own perspective (with the prompt YOU). They are then shown a single digit in the middle of the screen, followed immediately by a scene such as those shown in this figure. They are asked to respond "Yes" or "No" depending on whether the digit matches the number of dots in the picture. On "self" trials, participants must respond based on the number of dots they see in the picture. On "other" trials, they must decide whether the digit matches the number of dots the avatar sees. The classic result in this paradigm is that participants react more slowly in *inconsistent* scenes (as pictured on the right), in which participants can see a different number of dots than the avatar, than in *consistent* scenes (left), in which they and the avatar can see the same number of dots (Bukowski et al. 2015; Samson et al. 2010; Santiesteban et al. 2014). This *consistency* effect occurs both when participants are reporting the number of dots the avatar can see (i.e. reaction times are slowed by the participant's own perspective; this is called *egocentric interference*) and when participants report their own perspective (this is called *altercentric interference*).

cues participants to a specific location, an arrow provides a more general cue (Marotta et al. 2012).

A second series of modified DPT variants instead manipulates what the avatar appears able to see, using either barriers that block the dots from the avatar's field of view, or cartoon blindfolds or opaque goggles (Baker et al. 2016; Cole et al. 2016; Conway et al. 2017; Furlanetto et al. 2016). We call these "occlusion" tasks. The idea here is that if altercentric interference is driven by directional orienting, then it should occur whenever the number of dots the avatar faces is lower than the overall number of dots in the scene, even if the avatar cannot "see" the dots (e.g. due to an occluding barrier or other method of blinding). If the effect is instead driven by perspective-taking, altercentric interference should not appear when the avatar is blinded, since the avatar cannot "see" any of the dots in either consistent or inconsistent scenes. These tasks have produced contradictory results, with some finding effects supportive of the perspective-taking account (Baker et al. 2016; Furlanetto et al. 2016) and others supporting the directional orienting account (Cole et al. 2016; Conway et al. 2017; Langton 2018).

One possible explanation of these various contradictory results is that these experiments differ in whether participants are ever required to take the perspective of the avatar. Most of the experiments in the first, pioneering DPT study (Samson et al. 2010) required participants to answer based on their own perspective on some trials ("self" trials), and based on the avatar's perspective on others ("other" trials). We call these *explicit* tasks, because the avatar's perspective is explicitly relevant in these tasks. Explicit tasks can can establish the presence of both egocentric and altercentric interference: on "other" trials, explicit tasks may demonstrate *egocentric* interference, or slower judgements of the avatar's perspective due to interference from one's own perspective, since they are the only tasks that require participants to take the avatar's perspective. On "self" trials, explicit tasks may demonstrate *altercentric* interference, or slower judgements of one's own perspective due to interference from the avatar's perspective (see Figure 1).

This first DPT study (Samson et al. 2010) also included one task (Experiment 3) in which participants respond based only on their own perspective throughout the task. This experiment was motivated by concerns that mixing "self" and "other" trials may have cued participants to take the avatar's perspective even on trials where it was not relevant (i.e. on "self" trials). Participants were prompted with the cue "YOU" before every trial, and were instructed to ignore the central stimulus. We call this as an *implicit* task, because although it does not require participants to take the avatar's perspective as part of the task, it does overtly mention the avatar and its perspective – whether to instruct participants to ignore the avatar's perspective, as in Santiesteban et al. (2014), or to clarify for participants what the avatar can and cannot see, as in Cole et al. (2016). These instructions, along with the use of the word YOU as a cue on each trial, may still serve to prompt the participants to consider the avatar's perspective as relevant to the task, hence the label *implicit*. Implicit tasks are capable of establishing only altercentric interference, not egocentric interference; but altercentric interference is the effect that drives the claim of automatic/spontaneous perspective-taking, and so is the primary effect of interest in the DPT.

Note that the use of the terms "explicit" and "implicit" in this sense differ slightly from their use in the wider Theory of Mind literature, which distinguishes between explicit tasks that require a verbal response about another individual's mental states, and implicit tasks that infer the presence of the representation of another individual's mental states based on non-verbal responses (see e.g. San Juan and Astington 2017). Here, we are using the terms to refer to the task instructions and demands; that is, to describe whether participants are *explicitly* or *implicitly* required to take the perspective of the avatar throughout the task.

Occlusion tasks have generally opted to use either an explicit design throughout a battery of tasks, or an implicit design throughout. Those using explicit tasks have tended to find evidence consistent with the perspective-taking account (Baker et al. 2016; Furlanetto et al. 2016), while those using implicit tasks have tended to find evidence consistent with directional orienting (Cole et al. 2016; Conway et al. 2017; Langton 2018). One study has compared an explicit and implicit task, but this was done within subjects, in a substantially altered version of the DPT, making the findings difficult to interpret (Conway et al. 2017).

One further possibility is *uncued* tasks, which make no mention of perspective-taking in any of the information given to participants, have no requirement to take the avatar's perspective, and no trial-by-trial "YOU" cue that could implicitly contrast the participant's perspective with some other perspective. These tasks find no altercentric interference effect, unless there are further task modifications that draw additional attention to the avatar in some other way (for instance, having the avatar appear up to 600 ms before the dots in the scene) (Bukowski et al. 2015; Gardner et al. 2018b). (In the tasks that draw attention to the avatar in some way, results have been consistent with both the directional orienting and perspective-taking accounts, since these were not occlusion tasks. No existing uncued task attempts to discriminate between these.) In the *Supplementary Information* we describe a pilot study (uncued) that reports the same pattern of results.

In sum, apparently inconsistent results across variants of the DPT task may plausibly be due to differences in whether the perspective of the avatar – or other stimulus, such as an arrow – is made salient in one way or another, regardless of whether that perspective is strictly relevant for the task. This possibility prompts us to clearly distinguish between automatic and spontaneous cognitive processes, as described in the next section.

Implications for automaticity and spontaneity

Much of the experimental literature on the DPT is presented as informing the debate on "spontaneous perspective-taking" or "automatic perspective-taking". These terms are not often distinguished and sometimes used interchangeably. Few studies discuss exactly what spontaneity and/or automaticity entail. Where there is such discussion the most common approach is to say that for visual perspective-taking (or directional orienting) to be automatic or spontaneous, it should be purely stimulus-driven (Bukowski et al. 2015; Cole et al. 2016; Gardner et al. 2018b; Langton 2018). That is, it should occur reflexively and mandatorily on seeing the avatar, without any cues to participants to take the avatar's perspective, and without any need or motivation

on the part of the participants to consider the avatar's perspective relevant to the task (Cole et al. 2016; Gardner et al. 2018b; Langton 2018). Whether these conditions are appropriate can be disputed. For instance, some researchers have suggested that automaticity is best conceived of not as a binary, but rather as a matter of degree, in which features such as goal-directedness, intentionality, control, and purely stimulus-driven response each play a partial role in establishing whether a process is automatic (Moors and De Houwer 2006). Still, the more narrow definition of automatic as purely stimulus-driven is fairly widespread in the DPT literature.

We suggest that *automatic* and *spontaneous* cognitive processes should be clearly distinguished (see also Carruthers (2017); Westra (2017)). We consider automatic processes to be those that are reflexive and cannot be inhibited. In contrast, spontaneous processes are unconscious, involuntary, and rapid, but their operation is determined by intention, attention, or some other form of calibration. As an example of the difference, contrast seeing in colour, which is automatic, with seeing in focus, which is spontaneous: it occurs only as and when necessary, as determined by attention.

The varying empirical results reviewed above suggest two separate, but related, questions about visual perspective-taking:

(i) Does the altercentric interference effect found in the DPT provide evidence of visual perspective-taking or directional orienting?
(ii) Does the process driving altercentric interference (whether visual perspective-taking or directional orienting) arise automatically, spontaneously, or neither?

The current literature suggests that the principal effect in the DPT is moderated by top-down appraisal of the task context (Bukowski et al. 2015; Gardner et al. 2018b,a). In basic uncued tasks, with no awareness of the potential relevance of perspective-taking, there is no effect, while in uncued tasks when attention is drawn to the avatar in some way, there is an effect. In implicit tasks where there is minimal awareness of the presence of the avatars, there tends to be a directional orienting effect; visual perspective-taking effects occur only in explicit tasks, where there is a requirement to actively model the perspective of the avatars. In explicit tasks, perspective-taking is voluntary at certain points during the task, but is nonetheless involuntary on those trials where the avatar's perspective is irrelevant to the immediate question. This pattern suggests that perspective-taking is not automatic, but may be spontaneous – that is, occurring rapidly and involuntarily on individual trials where the avatar's perspective is irrelevant, but only in an overall task where perspective-taking is relevant.

We present five experiments (three preregistered novel experiments, two replications using different stimuli) testing the hypothesis that the varying results reported in the literature are a consequence of task design. We first contrast explicit, implicit and uncued versions of the DPT in a between-subjects design. Based on our reading of published results, we predicted that the explicit task would show an effect consistent with visual perspective-taking rather than directional orienting; that the implicit task would show directional orienting; and that the uncued task would show no effect. Findings matching these predictions would suggest a continuum of attention to the avatar's perspective, depending on motivation created by task context; and that both visual perspective-taking and directional orienting arise spontaneously but not

automatically. We then present a series of implicit tasks that attempt to establish the conditions under which an altercentric effect is found in the implicit condition.

Experiment 1: explicit, implicit and uncued

Materials and methods

We constructed a new set of stimuli using photographs of Lego figures, dubbed "Sally" and "Andrew" for ease of reference (Figure 2).[†] We did this in order to increase task complexity for a planned series of experiments (not reported here) using multiple avatars simultaneously. Unlike the cartoon avatar used in most DPTs to date, these scenes had the benefit of unambiguous depth in the third dimension, and solid black barriers were used to prevent any ambiguity in whether or not Lego figures were able to see through them. A variety of hiding places allowed balls (our equivalent of dots/discs) to be hidden from view of the Lego figures, even when placed in front of them. Specifically, the balls could appear in any of five positions: on a central table, visible to either figure; on either side of the table, at the feet of the Lego figure, and within view only of the figure on that side of the table; or on either external boundary of the scene, behind an external barrier, within view of neither figure. Each scene featured a single Lego character, either Sally or Andrew. Each figure could appear on either side of the screen, along with 0 to 4 balls and a maximum of two balls in any given location. The scenes were limited to 4 balls to allow for subitization: that is, rapid and accurate enumeration of low numbers of items. Trick and Pylyshyn (1994) find that reaction times remain low for subitization of four items or fewer.

This layout allowed for two different definitions of perspective consistency (Figure 3). Line-of-sight consistency captures the inconsistent/consistent distinction used in the original DPT: line-of-sight consistent scenes are those in which there are no balls occluded from the avatar's perspective; the avatar and the participant can see the same number of balls. Line-of-sight inconsistent scenes are those in which the participant can see balls that are hidden from the avatar. A second definition of consistency describes whether the balls are in the direction that the avatar faces, regardless of whether or not they are occluded: directionally consistent scenes are those in which all balls are placed in the direction the avatar faces, while directionally inconsistent scenes are those in which balls appear behind the avatar.

Scenes may therefore be consistent by both definitions (*Avatar sees*), inconsistent by both definitions (*Behind avatar*), or line-of-sight inconsistent but directionally consistent (*Avatar faces*). Line-of-sight consistent, directionally inconsistent scenes are not logically possible. Although previous research (Samson et al. 2010) controlled for the spatial layout of the room, confirming that the presence of the avatar and not merely the distance between the red dots was driving altercentric interference, it is possible that the greater complexity of our Lego scenes could introduce spatial artefacts. Specifically, scenes either have balls clustered entirely around the central table, or include balls on the periphery of the scene, outside the external walls. *Avatar*

[†]Materials, experiment code, data and analysis scripts for all experiments reported in this paper are available at https://osf.io/za3qd/.

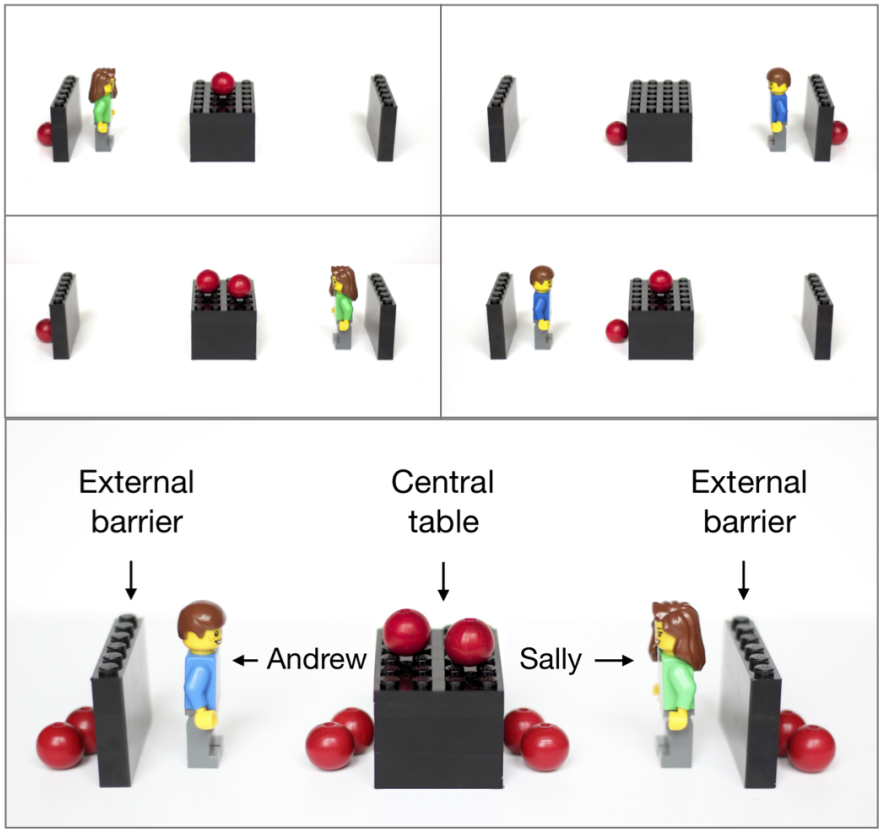


Figure 2. Adapted DPT stimuli using Lego figures. The upper four images show example scenes; note that each scene that participants saw featured a single avatar and a maximum of four balls. The lower image shows both potential placement positions for avatars (left or right of the central table) and all possible ball positions (5 possible positions, maximum of two balls in any one position).

sees scenes are necessarily central; *Behind avatar* scenes are necessarily peripheral. Some *Avatar faces* scenes have balls only around the central table, while some include peripheral balls. *Avatar faces* scenes are therefore categorised further into *Avatar faces (central)*, allowing a comparison with *Avatar sees* scenes that controls for the spatial distribution of balls from the centre of the scene; and *Avatar faces (peripheral)*, allowing a spatial distribution-controlled comparison with *Behind avatar* scenes.

Based on our review of the DPT literature above, we made the following specific predictions for altercentric interference (that is, from "self" trials only):

(i) Uncued, implicit and explicit tasks will all result in slower RTs for scenes with dots positioned behind the avatar, compared to dots positioned in front of, and visible to, the avatar (that is, *Behind avatar* vs *Avatar sees* trials). There

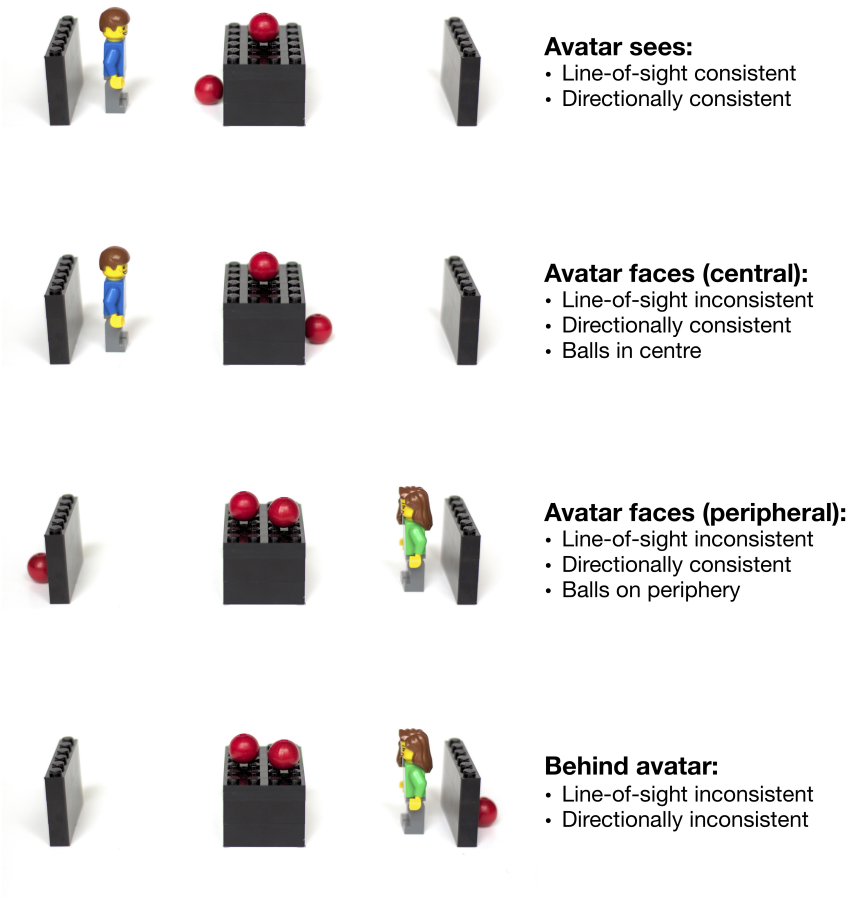


Figure 3. Example scenes from the four consistency conditions capturing the differences between avatar and participant perspectives, as well as spatial distribution.

are three possible explanations for this effect: the spatial distribution of the scene, directional orienting, or visual perspective-taking. Further comparisons will discriminate between these possibilities.

(ii) The explicit task will show visual perspective-taking rather than directional orienting, illustrated by slower RTs on *Avatar faces (central)* than *Avatar sees* trials; that is, a delay when some balls are not visible to the avatar, even when they are in the direction that the avatar is facing. In the implicit and uncued conditions, we predict no difference between *Avatar faces (central)* and *Avatar sees* trials, suggesting no visual perspective-taking in these conditions.

(iii) The implicit and explicit tasks will show directional orienting, illustrated by slower RTs on *Behind avatar* than *Avatar faces (peripheral)* trials. That is, trials

where all balls are in the direction the avatar is facing should be faster than those where balls are behind the avatar, suggesting directional orienting driving the *Behind avatar–Avatar sees* effect in the implicit task, and contributing to the effect in the explicit task. We expect that the uncued task will show no difference between *Behind avatar* and *Avatar faces (peripheral)* trials, suggesting that the *Behind avatar–Avatar sees* effect is driven purely by spatial distribution in this condition.

Preregistration. The experimental design and analysis was preregistered as part of the Open Science Framework's Preregistration Challenge; the timestamped plan is available at https://osf.io/5ey6d.

Participants. Simulations based on a pilot experiment (see SI Section 1) suggested that a sample size of 30 participants per condition would give substantially higher than 80% power at $\alpha = 0.05$ for the estimated effect sizes, for the within-subjects variables of interest. Ninety participants were recruited through the University of Edinburgh Student and Graduate Employment Service, and assigned randomly to the three between-subjects conditions: explicit, implicit and uncued. They were compensated £4 for their participation, which lasted approximately 30 minutes. Data were excluded from two participants whose tasks were interrupted by computer failure, and replaced by data from two new participants.[‡] Participants gave written consent, including consent for anonymised data to be shared publicly. Ethical approval was granted by the University of Edinburgh's School of Philosophy, Psychology and Language Sciences Research Ethics Committee (PPLSREC), reference number 109-1718/1.

Procedure. On each trial, participants saw a fixation cross, followed by a one-word instruction, followed by a digit (0–4) presented for 750 ms, finally followed by a Lego scene accompanied by a prompt for a response. Figure 4 shows example trial sequences. In the explicit condition, participants were told that their task was to judge whether the digit they saw on each trial matched the number of balls that could be seen in the following picture. If they saw the word YOU before the trial ("Self" trials), they should answer based on how many balls they could see, and if they saw the word HE or SHE before the trial ("Other" trials), they should answer based on how many balls the Lego figure could see. In the implicit condition, participants were instructed to ignore what the Lego figure could see, and answer based only on what they could see. They were told that the word YOU would appear before each trial in order to remind them to answer based on their own perspective. In the uncued condition, participants were told that their task was to judge whether the digit matched the number of balls in the picture, with no mention of the Lego figure. The word READY? appeared before each trial, in order to make the trial length identical across conditions.

Participants completed a short training session explaining the task, followed by 32 practice trials (each followed by feedback informing the participant whether their answer had been correct or incorrect), and then the main task, divided into four blocks with self-paced breaks between blocks. On each trial, participants were presented with

[‡]Note that this is a technical deviation from the preregistration, in which we did not explicitly state that we would replace such participants.

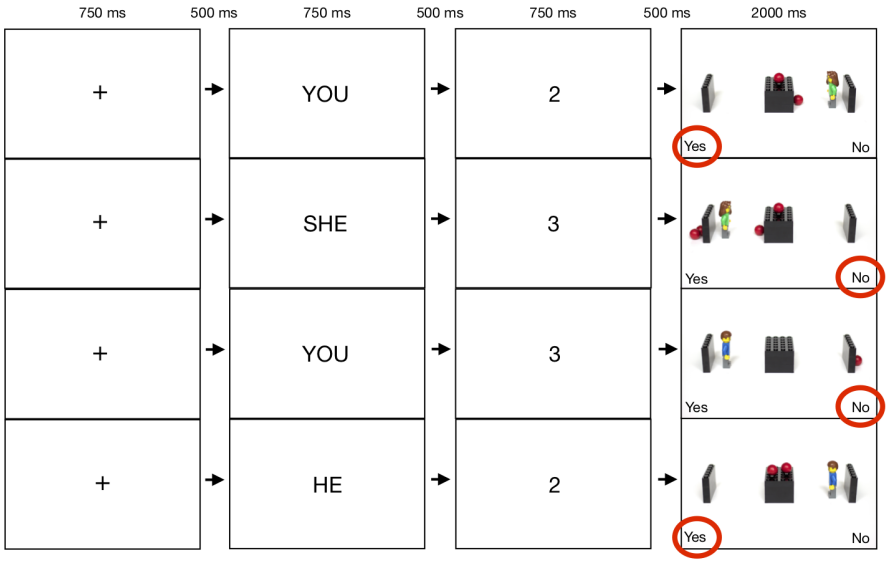


Figure 4. An illustration of the trial procedure in the explicit condition, with correct answers highlighted.

the cue word (YOU/HE/SHE/READY?, depending on condition) for 750 ms, followed by a fixation cross for 750 ms, and finally a digit between 0 and 4 for 750 ms, before the Lego scene appeared with the words "Yes" and "No" in the bottom corners of the screen. A two-button button box was used to respond, with participants instructed to press the Yes-side button for yes and the No-side button for no. The "Yes" and "No" labels were presented on the screen to facilitate exact replication between tasks regardless of input equipment. The sides of these prompts were counterbalanced between participants, with half of the participants seeing "No" on the bottom left-hand corner of the screen throughout the task, and the other half seeing it on the bottom right-hand corner. This counterbalancing was done to avoid left-to-right reading bias possibly favouring the left-hand prompt, and majority human left hemispheric dominance possibly favouring the right-hand prompt. Scenes timed out within 2000 ms if no response was given, and the task moved on to the following trial.

The manipulated within-subjects variable of interest was the consistency between the avatar's perspective and the participant's perspective. For each participant, there were 64 trials in each of the four consistency conditions (*Avatar sees*, *Avatar faces (central)*, *Avatar faces (peripheral)*, and *Behind avatar*). Additionally, a range of other constraints were followed, balancing which avatar appeared and on which side of the scene, the number of scenes with each possible number of balls, the number of yes vs no answers, and in the explicit condition, Self vs Other trials (see SI section 2).

The experiment was implemented using PsychoPy (Peirce 2010).

Results

This design allowed the predictions detailed above to be tested using a series of mixed effects models. All analyses reported below are in accordance with the preregistered analysis plan, unless otherwise noted.

We removed training trials, filler trials (those with zero balls), and timed-out trials (0.76%, $n = 176$); as per Whelan (2008), trials in which the response RT was lower than 100 ms were also removed, on the assumption that these trials could not be genuine responses to the stimuli (0.02%, $n = 5$). No trimming was conducted on higher reaction times, given the imposed cut-off of 2000 ms on all trials. Although Samson et al. (2010) and many subsequent DPT variants analyse only "Yes" trials on the basis that "No" trials may be easier to respond to, Santiesteban et al. (2014) found no difference between "Yes" and "No" responses. We therefore have not removed data from "No" trials and do not include this as a variable in our analyses. Because reaction time data deviates from the normal distribution, models used a log transform of reaction time as the dependent variable, in order to reduce skewing of the data and better conform to the assumptions of the model (Baayen and Milin 2010).

A binomial logistic regression analysis of error rates in a pilot task (reported in the *Supplementary Information*) failed to converge, presumably due to a lack of data, since error rates in the DPT are extremely low. We therefore removed trials with erroneous responses (4.2%, $n = 959$), but did not analyse them due to a lack of statistical power.

Because we were interested in altercentric interference, all "Other" trials were removed in the explicit condition (recall that the definition of implicit and uncued conditions is that there are no "Other" trials).[§] This means that in all conditions, we are looking only at participants' responses where they are evaluating whether the digit they were presented with matches the number of balls *they* can see. Any interference effects will therefore be altercentric interference, i.e. due to inability to suppress the avatar's perspective when that perspective is irrelevant on the trial at hand.

The explicit, implicit and uncued tasks were the same length to avoid differing fatigue effects between conditions, which halved the number of trials available for analysis in the explicit task: 3607 explicit trials vs. 7235 implicit and 7396 uncued.[¶] Given that all analyses were within-subjects in a particular condition and that the power analysis showed sufficient statistical power for this number of trials in the explicit condition, we see no reason that this could have accounted for any differences between conditions. We had no theoretically motivated predictions for Other trials, and these trials were therefore not analysed in order to limit researcher degrees of freedom in the analysis (see Simmons et al. (2011) for discussion of the problems associated with researcher degrees of freedom).

[§]This step was included in the preregistered analysis script, and is necessitated by the experiment design, but was erroneously omitted from the preregistration free-text description of the analysis.

[¶]A reviewer suggested an exploratory analysis using only the first half of the trials from the implicit and uncued tasks, in order to confirm whether diminishing effects in the implicit and uncued tasks may explain the results reported. This analysis, using the trials only from the first two blocks of the implicit and uncued tasks, produces the same pattern of results as the preregistered analysis reported here, i.e. the same comparisons produced similarly significant or null results. One minor exception is noted in the results of Model 3. The script for this exploratory analysis is available at osf.io/za3qd/.

We used `lme4` (Bates et al. 2015) and `afex` (Singmann et al. 2017) to perform a series of mixed effects regression analyses on the log-transformed reaction times (logRT). Mixed-effects models were used rather than the ANOVA used in previous experiments to avoid the necessity of averaging across observations for each participant, and to account for random effects – that is, the variance associated with different images as well as different participants. We used the standard $p < .05$ criterion for determining where effects were significant, with p-values obtained using model comparison (likelihood ratio tests) using the `mixed()` function in the `afex` package (Singmann et al. 2017) in R ({R Core Team} 2015).

Model 1: Are Behind avatar scenes slower than Avatar sees scenes? We predicted that RTs in *Avatar sees* scenes (where the avatar's perspective matched the participant's) and in *Behind avatar* scenes (where the avatar's perspective mismatches the participant's, under both line-of-sight or directional accounts) should differ (specifically, RTs in *Avatar sees* scenes should be faster) in all three tasks (explicit, implicit and uncued), although possibly for different reasons, to be unpicked in subsequent analyses. In order to test this prediction, we ran an analysis on the data from *Avatar sees* and *Behind avatar* trials. Consistency and Condition (explicit vs implicit vs uncued) were sum-coded and entered as fixed effects, with interaction term, into the model. The sum coding for condition resulted in comparisons of explicit vs implicit, and implicit vs uncued. As differences in overall RT between the three conditions were not relevant to our predictions and had no theoretically-motivated hypotheses about these differences, the results of these slopes are not reported. Random intercepts for images and participants were specified, as well as by-participant random slopes for the effect of consistency.[‖]

The model (Table 1) showed an effect of Consistency, suggesting that *Behind avatar* trials were approximately 44.22 ms slower on average than *Avatar sees* trials. There was no interaction between Condition and Consistency, implying that all three conditions showed the same effect, with a 59.89 ms difference in the explicit condition, 38.93 ms in the implicit condition, and 35.66 ms in the uncued condition (Figure 5).

In all conditions, then, *Avatar sees* trials were associated with faster RTs than *Behind avatar* trials, matching our prediction. However, the cause of this effect (visual perspective-taking, directional orienting, or spatial distribution) is unclear.

Model 2: Is there a mentalizing effect in the explicit condition? We limited our data to *Avatar sees* and *Avatar faces (central)* trials – recall that in *Avatar faces (central)* trials, all balls in the scene are located centrally, but the participant and the avatar have distinct line-of-sight perspectives, i.e. some balls are 'hidden' from the avatar behind the central table. Otherwise, the model was identical to Model 1. The model (Table 2) showed no significant effect of Consistency, but a significant interaction between Condition and Consistency. Planned pairwise comparisons showed that *Avatar faces (central)* trials were, on average, 27.79 ms slower than *Avatar sees* trials in the explicit condition, but showed no significant difference in the implicit or uncued conditions (Figure 6). This

[‖]Model syntax:
`logRT ~ Condition*Consistency + (1+Consistency|Participant) + (1|Image)`

Table 1. Results of Experiment 1, Model 1: *Avatar sees* vs *Behind avatar*

Model	Slope	β	SE	χ^2	df	*p*
Main model	Consistency	0.039	0.004	48.49	1	$< .001^{***}$
	Consistency x Condition (implicit vs explicit)	0.007	0.005	2.36	2	.31
	Consistency x Condition (implicit vs uncued)	0.005	0.004			
Planned comparisons	Consistency (explicit)	0.044	0.008	22.62	1	$< .001^{***}$
	Consistency (implicit)	0.034	0.005	32.71	1	$< .001^{***}$
	Consistency (uncued)	0.032	0.006	24.70	1	$< .001^{***}$

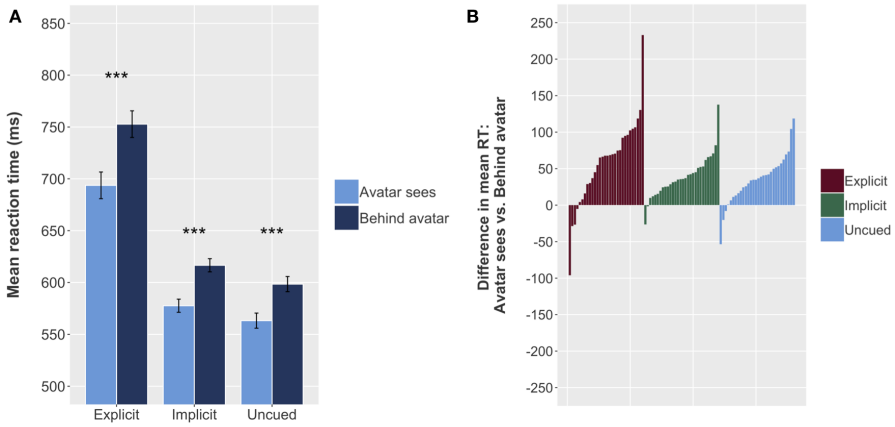


Figure 5. Effects of Experiment 1, Model 1: *Avatar sees* vs *Behind avatar*. (A) Mean RT for *Behind avatar* and *Avatar sees* conditions, for explicit, implicit and uncued tasks; error bars indicate 95% CIs on the mean of the by-participant means, and significance annotations on the plots reflect the planned comparisons showing the effect of consistency for each condition. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT; lines extending above 0 on the y-axis indicate that the participant was slower in *Behind avatar* than in *Avatar sees* trials (i.e. exhibited an altercentric interference-like effect), while lines extending below 0 indicate that the participant was slower in *Avatar sees* than in *Behind avatar* trials. Mean reaction time is higher (i.e. participants respond more slowly) for *Behind avatar* trials in all three conditions (A); a substantial majority of participants in all three conditions show this effect (B).

matches our prediction, and suggests visual perspective-taking in the explicit condition, and either a directional orienting or a spatial-distribution effect underlying the results for the implicit and uncued conditions in Model 1.

Table 2. Results of Experiment 1, Model 2: *Avatar sees* vs *Avatar faces*

Model	Slope	β	SE	χ^2	df	*p*
Main model	Consistency	0.005	0.004	1.53	1	.22
	Consistency x Condition (implicit vs explicit)	0.015	0.005	11.32	2	.003***
	Consistency x Condition (implicit vs uncued)	0.008	0.004			
Planned comparisons	Consistency (explicit)	0.021	0.009	4.93	1	.03*
	Consistency (implicit)	−0.005	0.004	1.10	1	.29
	Consistency (uncued)	−0.003	0.005	0.48	1	.49

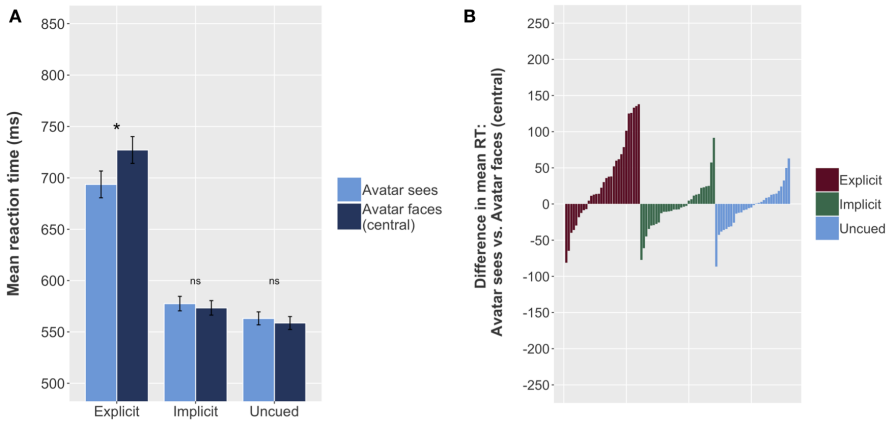


Figure 6. Effects of Experiment 1, Model 2: *Avatar sees* vs *Avatar faces*. (A) Mean RT for *Avatar faces (central)* and *Avatar sees* conditions, for explicit, implicit and uncued conditions; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Avatar sees* RT and mean *Avatar faces (central)* RT. Mean reaction time is higher (i.e. participants respond more slowly) for *Avatar faces (central)* trials in the explicit condition, but not in the implicit or uncued conditions (A); a substantial majority of participants in the explicit condition, but not in the implicit or uncued conditions, show this effect (B).

Model 3: Is there a directional orienting effect in the implicit condition? We limited our data to *Avatar faces (peripheral)* and *Behind avatar* trials – recall that in *Avatar faces (peripheral)* trials the participant and the avatar have distinct line-of-sight perspectives, i.e. some balls are 'hidden' from the avatar in a peripheral position, in

the direction that the avatar is facing but behind one of the outer barriers; in *Behind avatar* trials some balls are 'hidden' behind the avatar, again in a peripheral position. The model (Table 3, model structure and coding as per previous analyses) showed no effect of consistency and no interaction between condition and consistency.** Planned pairwise comparisons showed a significant effect for explicit but not implicit or uncued conditions. Note however that given the omnibus model showed no interaction, the significant effect (24.24 ms) in the model analysing the explicit condition only should be treated with caution (Figure 7).

Table 3. Results of Experiment 1, Model 3: *Avatar faces (peripheral)* vs *Behind avatar*

Model	Slope	β	SE	χ^2	df	*p*
Main model	Consistency	0.008	0.004	3.92	1	.05
	Consistency x Condition (implicit vs explicit)	0.009	0.005	3.90	2	.14
	Consistency x Condition (implicit vs uncued)	−0.005	0.004			
Planned comparisons	Consistency (explicit)	0.018	0.007	5.45	1	.02*
	Consistency (implicit)	0.004	0.006	0.45	1	.50
	Consistency (uncued)	0.004	0.005	0.58	1	.44

There is therefore (somewhat tentative) evidence matching our predictions for the explicit condition (where we expected a directional orienting component to visual perspective-taking, here indicated by participants responding faster when balls were in the direction the avatar was facing, even though they were occluded from the avatar's line of sight). These results also match our prediction for the uncued condition, where we predicted no effect of the avatar, and a *Behind avatar–Avatar sees* altercentric interference effect (see Model 1) driven entirely by central vs peripheral distributions of the balls. However, we do not find evidence matching our prediction for the implicit condition, where we predicted altercentric interference in this model, driven by directional orienting. Rather, our results suggest that our implicit and uncued conditions behave similarly, with the only effects we see being driven by the spatial distribution of the balls, with slower responses to scenes featuring balls in the periphery of the scene.

Discussion

These results support our hypothesis that the requirement to take the avatar's perspective on some trials results in visual perspective-taking; and that differences

**The exploratory analysis mentioned above using trials from only the first two blocks in the implicit and uncued conditions showed a consistent interaction between condition and consistency, $p = 0.03$.

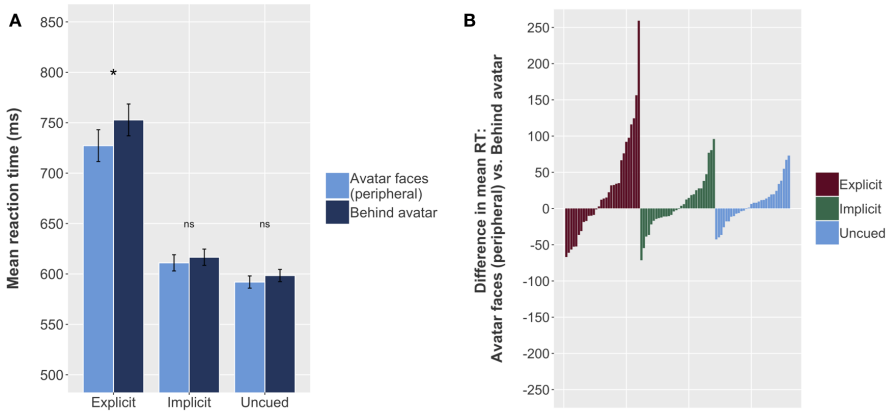


Figure 7. Effects of Experiment 1, Model 3: *Avatar faces (peripheral)* vs *Behind avatar*. (A) Mean RT for *Behind avatar* and *Avatar faces (peripheral)* conditions, for explicit, implicit and uncued conditions; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar faces (peripheral)* RT. Mean reaction time is higher (i.e. participants respond more slowly) for *Avatar faces (peripheral)* trials in the explicit condition, but not in the implicit or uncued conditions (A); a small majority of participants in the explicit condition, but not in the implicit or uncued conditions, show this effect (B).

in task design or framing explain apparently conflicting results in the literature. We can manipulate the presence/absence of an altercentric interference effect by switching between an explicit task and implicit or uncued tasks.

In the explicit task, *Avatar sees* trials were 59.89 ms faster than *Behind avatar* trials, suggesting a spatial, perspective-taking, or directional orienting effect, or some combination of the three; *Avatar sees* trials were 27.79 ms faster than *Avatar faces (central)* trials, suggesting perspective-taking; and *Avatar faces (peripheral)* trials were 24.24 ms faster than *Behind avatar* trials, suggesting directional orienting. The considerably larger effect in *Behind avatar* vs *Avatar sees* suggests that the processes may be summative; that is, with both the distribution of balls from the centre of the scene and the avatar's perspective causing individual delays that result in a larger overall delay. The evidence for directional orienting (although this evidence is tentative, given the lack of omnibus effect in this model) suggests that directional orienting may play a role in perspective-taking, or otherwise operate in tandem with it, perhaps as a first visual sweep of a scene. The precise interaction of these varying effects would be a useful subject for future research.

The results are coherent with previous research using explicit and uncued tasks, but conflict with several studies that find altercentric interference in implicit tasks (Samson et al. 2010; Santiesteban et al. 2014; Langton 2018), likely driven by directional orienting (Cole et al. 2016; Conway et al. 2017).

A potentially important difference between our Lego stimuli and standard DPT stimuli is the positioning of the avatar. Previous implicit tasks (Samson et al. 2010;

Santiesteban et al. 2014; Cole et al. 2016; Conway et al. 2017; Langton 2018) have placed the avatar in the centre of the screen, preceded by a fixation cross and trial-by-trial instructions in the centre of the screen. Our stimuli position the avatar off-centre, preceded by the fixation cross and trial-by-trial instructions in the centre of the screen. Given the literature suggesting that additional attention drawn to the avatar induces an altercentric effect even on uncued tasks (Bukowski et al. 2015; Gardner et al. 2018b), it is possible that previous implicit tasks have drawn additional attention to the avatar through the placement of the fixation cross and instructions over the spot where the avatar will appear (see e.g. Bukowski et al. (2015)).

We therefore conducted a second preregistered implicit task, identical to the implicit condition in Experiment 1 but with the fixation cross and trial-by-trial instructions (i.e. the text "YOU" and the digit to be confirmed) placed directly over the point on the screen where the avatar will appear. We predicted that we would find the expected *Avatar faces (peripheral)* vs *Behind avatar* altercentric interference in this condition. This would suggest that attention must be drawn directly to the avatar on a trial-by-trial basis in order to induce directional orienting, implying that neither visual perspective-taking nor directional orienting is automatic; rather, they appear in response to ongoing cues regarding the avatar's relevance to the task.

Experiment 2: salience of avatars

Materials and methods

The same stimuli were used as for Experiment 1.

Preregistration. The experimental design and analysis was preregistered as part of the Open Science Framework's Preregistration Challenge; the timestamped plan is available at `https://osf.io/dk86n`.

Participants. Simulations based on Experiment 1 suggested that a sample size of 30 participants per condition would give substantially higher than 80% power at $\alpha = 0.05$ for the estimated effect sizes, for the within-subjects variables of interest. Thirty further participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 30 minutes. Data were excluded from one participant whose task was interrupted by computer failure, and replaced by data from a new participant. Participants gave written consent, including consent for anonymised data to be shared publicly. Ethical approval was granted by the University of Edinburgh's PPLSREC, reference number 188-1718/1.

Procedure. This task used the same procedure and task design as the implicit condition in Experiment 1. Fixation crosses and pre-trial instructions (the appearance of the word YOU and the digit between 0 and 4) were changed to appear centred over the position in which the face of the Lego character would appear in the following scene, rather than appearing centrally on the screen.

Table 4. Results of Experiment 2

Slope	β	SE	χ^2	df	p
Behind avatar vs *Avatar sees*	0.028	0.005	23.72	1	$< .001^{***}$
Avatar faces (central) vs *Avatar sees*	−0.001	0.005	0.07	1	.80
Behind avatar vs *Avatar faces (peripheral)*	0.001	0.006	0.04	1	.84

Results

We applied the data exclusions and transformations described in Experiment 1, removing timed-out trials (0.49%, $n = 38$) and erroneous responses (3.36%, $n = 257$). There were no responses below 100 ms.

Following our preregistered analysis plan, data limited to the three relevant comparisons (*Behind avatar* vs *Avatar sees*, *Avatar faces (central)* vs *Avatar sees*, and *Avatar faces (peripheral)* vs *Behind avatar*) were modelled using three models identical to the pairwise comparisons for the implicit condition in Experiment 1.

As in Experiment 1, and as predicted, the models showed a significant difference between *Avatar sees* and *Behind avatar* (35.96 ms), and no significant difference between *Avatar faces (central)* and *Avatar sees* trials (−1.55 ms, see Table 4). However, contrary to our predictions, there was also no significant difference between *Avatar faces (peripheral)* and *Behind avatar* trials, at 1.43 ms (Figure 8). This suggests that there was no directional orienting effect in this task, and that the difference between *Avatar sees* and *Behind avatar* trials was driven by the spatial distribution of the balls.

Discussion

These results do not support the hypothesis that directional orienting played any role in this implicit task. These results continue to conflict with findings of consistency effects in implicit tasks (Samson et al. 2010; Santiesteban et al. 2014; Langton 2018; Cole et al. 2016; Conway et al. 2017).

One possible explanation for this could be the complexity of the scene. The original DPT used a simple scene consisting only of the avatar in a room, with an array of dots. Occlusion tasks have used a single avatar that appeared in a consistent position, with up to three balls and one or two barriers (Baker et al. 2016; Cole et al. 2016) or another method of blinding that added a single element to the existing scene, such as goggles or a telescope (Conway et al. 2017; Furlanetto et al. 2016). It may be that the Lego stimuli, with three barriers, two possible avatars appearing in two different places, and up to four balls, increased the scene complexity to the extent that participants' strategies to complete the task changed substantially. That is, it may be the case that participants were best able to respond quickly and accurately to each trial by ignoring the perspective of the on-screen character – a strategy that would not be possible in an explicit task (explaining the results in Experiment 1) but would be possible in implicit and uncued tasks.

To explore the possibility of scene complexity driving the null results in these implicit tasks, we conducted a further preregistered implicit task, simplifying the Lego

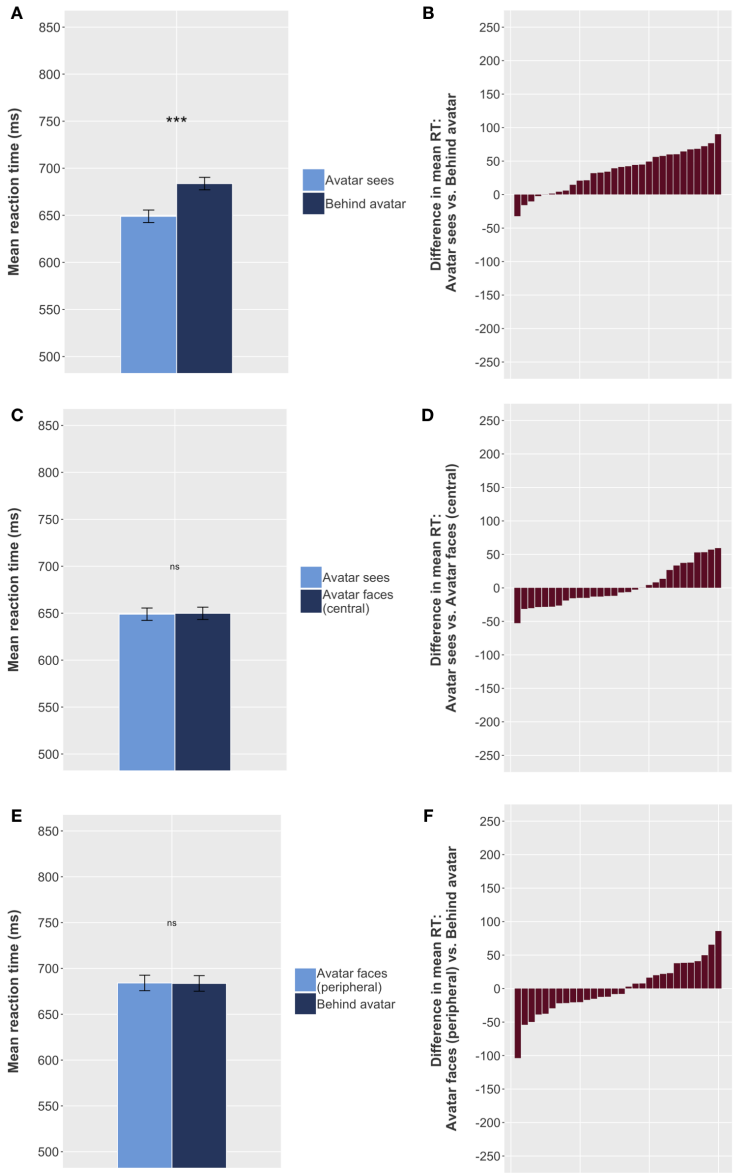


Figure 8. Results of Experiment 2. (A) Mean RT for *Behind avatar* and *Avatar sees* conditions; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT. Mean reaction time is higher (i.e. participants respond more slowly) for *Behind avatar* trials (A); a majority of participants show this effect (B). However, there is no difference in RT between *Avatar sees* and *Avatar faces (central)* (C, D) or *Behind avatar* and *Avatar faces (peripheral)* (E, F).

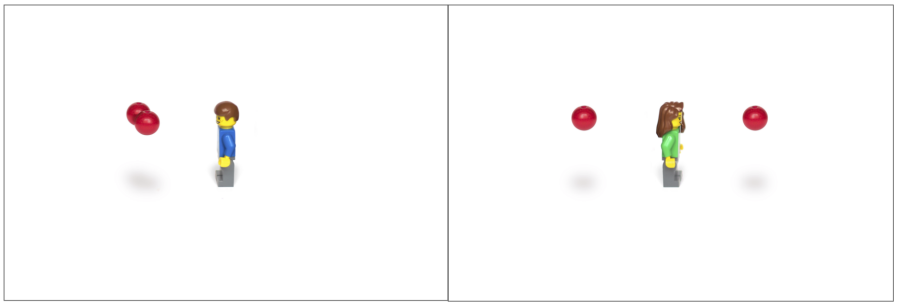


Figure 9. Lego stimuli adapted to match original DPT scene layout. Each scene consists of a single avatar and up to two balls, which can appear in front of or behind the avatar.

stimuli to scenes equivalent to those in the original DPT. These scenes consisted of a central figure, with balls level with the character's gaze, positioned either in front of or behind the character. Based on our reading of the extant literature, we predicted that scene complexity would explain the lack of altercentric effect on implicit tasks in Experiments 1 and 2, i.e. that there would be an altercentric effect with simplified stimuli.

Because these simplified stimuli do not incorporate any barriers that distinguish between *Avatar sees* and *Avatar faces*, they are not be able to provide evidence for whether any altercentric effect found in this task is better explained by perspective-taking or by directional orienting. However, the results of this task should help to explain the unexpected null results for directional orienting in the implicit task in Experiment 1, and in Experiment 2.

Experiment 3: Reducing the visual complexity of the scene

Materials and methods

The images used in Experiment 1 and 2 were digitally edited to match the layout of the original DPT stimuli (Figure 9). Each Lego character appeared centrally on the screen, facing either left or right, with up to two balls in each scene. The balls, which floated at the height of the gaze of the Lego character, could appear in front of the character, behind it, or both in front and behind. As in Experiment 2, participants were instructed to ignore the perspective of the Lego character, and the word YOU appeared before each trial. The fixation cross and pre-trial instructions appeared over the position where the face of the Lego character would appear.

Preregistration. The experimental design and analysis was preregistered as part of the Open Science Framework's Preregistration Challenge; the timestamped plan is available at https://osf.io/hr98w.

Participants. Sample size calculation was based on the same simulation method as Experiment 2. Thirty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their

Table 5. Results of Experiment 3

Slope	β	SE	χ^2	df	p
Behind avatar vs *Avatar sees*	−0.002	0.003	0.43	1	.51

participation, which lasted approximately 30 minutes. Participants gave written consent, including consent for anonymised data to be shared publicly. Ethical approval was granted by the University of Edinburgh's PPLSREC, reference number 188-1718/1.

Procedure. The procedure for Experiment 2 was used, with some differences. Participants completed 16 practice trials, followed by 192 test trials: 96 in which the avatar could see the same number of balls as the participant (*Avatar sees*) and 96 in which at least one ball was concealed behind the avatar (*Behind avatar*). Up to two balls appeared in any given scene. Avatar, yes/no response, and number of balls were balanced across trials (see SI Section 3).

Results

We applied the data exclusions and transformations described in Experiment 1, removing timed-out trials (0.30%, $n = 17$), erroneous responses (2.25%, n = 129), and the single trial with a response below 100 ms (0.02%).

Following our preregistered analysis plan, a mixed effects regression was used to compare the log-transformed reaction times for *Avatar sees* trials compared to *Behind avatar* trials. Consistency was sum-coded and entered as a fixed effect, and random intercepts for images and participants were specified, as well as by-participant random slopes for the effect of Consistency. Contrary to our prediction, the model showed no difference between *Avatar faces* and *Behind avatar* trials, at an estimated −2.60 ms (Table 5, Figure 10). This suggests (alongside the results of Experiments 1 and 2) an absence of any directional orienting in our implicit DPT. This is contrary to the findings of several existing studies (Samson et al. 2010; Santiesteban et al. 2014; Langton 2018; Cole et al. 2016; Conway et al. 2017).

Discussion

This task found no evidence of difference in RT between *Behind avatar* and *Avatar sees* in an implicit task, contrasting with the results of Experiments 1 and 2. This contrast may be explained by differences in the spatial distribution of the balls: in Experiments 1 and 2, *Avatar sees* trials all had balls clustered in the centre of the screen, while *Behind avatar* trials had balls on the periphery of the scene. In Experiment 3, these two conditions had balls evenly placed from the centre of the screen. The lack of effect in Experiment 3, with balls evenly distributed from the centre of the scene in these two conditions, therefore contributes to the evidence that this effect was driven by spatial distribution in Experiments 1 and 2.

The difference between the null result in Experiment 3 and the altercentric effect found in several implicit tasks (Samson et al. 2010; Santiesteban et al. 2014; Langton

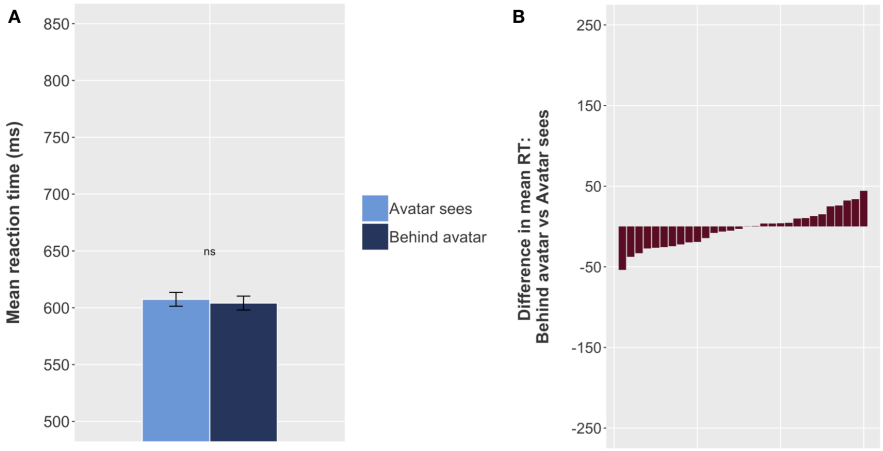


Figure 10. Results of Experiment 3. (A) Mean RT for *Behind avatar* and *Avatar sees* conditions; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT. Mean reaction time is not significantly different between the two conditions.

2018; Cole et al. 2016; Conway et al. 2017) raises the possibility that there is an important difference between the Lego stimuli and the avatars used in previous tasks. While we find this surprising, it may be due to unanticipated differences in the willingness of participants to accept Lego avatars vs cartoon avatars as having a perspective. Given that many participants are likely to have interacted with Lego characters as objects, but all would encounter the avatars for the first time in the context of the experiment, one possibility is that participants are more inclined to consider the Lego characters as objects but the cartoon-like avatars in the standard DPT stimuli as agents. The greater realism of the standard avatars may also contribute to a heightened perception of agency. We therefore ran a second simplified task using the original DPT stimuli, otherwise identical to Experiment 3. Because this was simply a replication of Experiment 3 using different stimuli, it was not preregistered separately, as all other details of the preregistration were the same.

Experiment 4: original stimuli

Materials and methods

The materials and methods for Experiment 3 were re-used, with original DPT stimuli instead of Lego stimuli. The images were sized so that the on-screen characters were the same height, and the characters' heads in the same position on the screen, as the Lego characters in Experiment 3.

Participants. Thirty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 30 minutes. Participants gave written

Table 6. Results of Experiment 4

Slope	β	SE	χ^2	df	*p*
Behind avatar vs *Avatar sees*	0.010	0.004	7.05	1	.008**

Table 7. Lego vs original stimuli

Slope	β	SE	χ^2	df	*p*
Original vs Lego	−0.011	0.027	0.18	1	.67
Behind avatar vs *Avatar sees*	0.004	0.003	2.34	1	.13
Interaction	0.006	0.003	5.83	1	.02*

consent, including consent for anonymised data to be shared publicly. Ethical approval was granted by the University of Edinburgh's PPLSREC, reference number 188-1718/1.

Procedure. This task used the same procedure and task design as Experiment 3.

Results

We applied the data exclusions and transformations described in Experiment 1, removing timed-out trials (0.24%, n = 14) and erroneous responses (2.98%, n = 171). There were no responses below 100 ms.

The data were analysed using a model identical to that used in Experiment 3. The results showed a significant difference between *Behind avatar* and *Avatar sees* trials, at 11.30 ms (Table 6). Note that this is a substantially smaller effect than other implicit tasks using these stimuli: 21 ms (Samson et al. (2010), Experiment 3); 35.4 ms (Santiesteban et al. (2014), Experiment 2); approximately 40 ms (Cole et al. 2016); 35 ms (Conway et al. (2017), Experiment 1). We conducted a further exploratory model comparing reaction times across the two experiments, with Consistency and Stimulus entered as fixed effects (with interaction term) and the same random effects structure as the basic model. This model revealed a significant Consistency x Stimulus interaction, providing further evidence of a consistency effect with the original stimuli, but not with the Lego stimuli (Table 7, Figure 11).

Discussion

These results suggest that, remarkably, the stimuli themselves play a role in producing an altercentric effect. The lack of an effect in the implicit tasks in Experiments 1, 2 and 3 appears to be due to some difference between the Lego stimuli and the original stimuli, suggesting that the Lego stimuli do not result in either directional orienting or perspective-taking without additional direction to take the character's perspective. It is possible that there are features of the scenes other than the avatars themselves driving this difference (for instance, the brightness of the colours; the overlap of balls in the Lego scenes compared to the spacing of the spacing of discs in the original stimuli; or the lack of a blue background room in the Lego scenes). A reviewer suggests that an alternative explanation is a difference between the directional features of Lego and

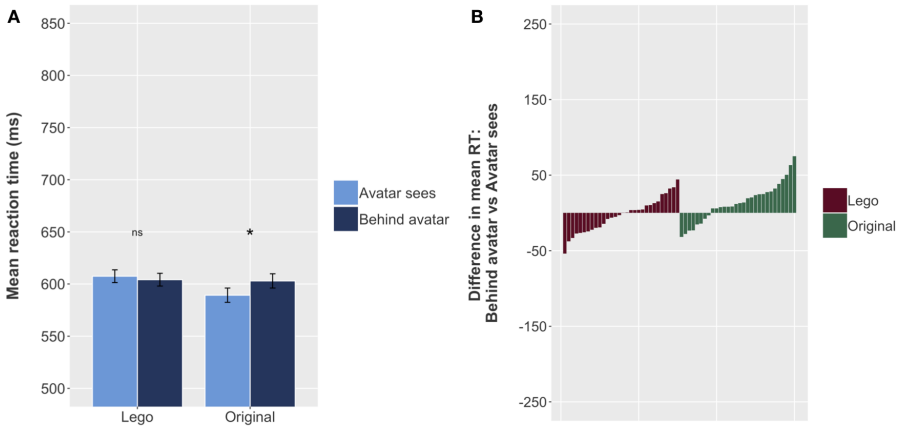


Figure 11. Comparison of effects in Experiments 3 and 4. (A) Mean RT for *Behind avatar* and *Avatar sees* conditions for both Lego and original stimuli; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT for both Lego and original stimuli. Mean reaction time is higher (i.e. participants respond more slowly) for *Behind avatar* trials for the original stimuli only (A); a majority of participants in this condition show this effect (B).

cartoon avatars. That is, the original avatars may provide more cues for the front and back of the body compared to the Lego avatars: they have torsos with a clear front and back shape, and faces with humanoid profiles, compared to the flat faces and body shape of the Lego pieces. Since non-humanoid stimuli have resulted in an altercentric effect (Santiesteban et al. 2014) in an implicit task, the directional cueing properties of the stimuli may play an important role.

Yet another explanation could be differences in attribution of agency to the Lego avatars compared to the cartoon avatars. This could be due to participants' experience of Lego characters as non-agentive objects in the real world, although Lego figures have been shown to be processed as animate in at least some circumstances (LaPointe et al. 2016); or it may be due to intrinsic properties of the images – that is, the greater realism of the cartoon avatars, with near-human proportions, body shape, and facial projections.

Altercentric interference appears to be a robust effect in a wide range of simple DPT variants, and has even been found in more complex scene layouts with non-standard avatars (Baker et al. 2016; Mattan et al. 2015). The unexpected lack of altercentric interference in Experiments 1, 2 and 3 can nonetheless be reconciled with the wider literature. The DPT variants that have used non-standard avatars have been explicit, and our Experiment 1 using Lego figures suggests that an explicit task may be sufficient to drive perspective-taking. Implicit tasks make up a limited sub-section of the DPT literature, and all use the standard avatar, or the standard avatar with minor modifications such as a blindfold or barrier (Cole et al. 2016; Conway et al.

2017; Samson et al. 2010; Santiesteban et al. 2014). There are two notable exceptions. First, Langton (2018) uses photographs of people in an implicit occlusion task, finding results consistent with directional orienting. This is coherent with both explanations discussed above; that is, that a photograph of a human would provide greater directional cues or clearer evidence of agentiveness than Lego figures in the same way that humanoid avatars would. Langton (2018) also uses live human experimenters in an uncued task that has a substantial delay between the orientation of the experimenter and the appearance of dots; as this is analogous to other uncued tasks that manipulate stimulus onset asynchrony (Gardner et al. 2018b; Bukowski et al. 2015), the finding of directional orienting in this task is not surprising. Second, Santiesteban et al. (2014) find an altercentric effect on an implicit task using arrows as stimuli; as discussed above, this is consistent with the explanation that sufficient directional cueing in a stimulus may be sufficient to trigger directional orienting.

These results contribute to the evidence suggesting that, while the altercentric effect may be widely replicated, it is nonetheless surprisingly sensitive to small differences in task design that prompt attention to the avatar and the relevance of its perspective. Prompts hinting at the relevance of certain kinds of agentive stimuli, such as discussion of the avatar's perspective and the inclusion of the YOU cue on every trial, may produce the altercentric effect in implicit tasks (Samson et al. 2010; Santiesteban et al. 2014; Langton 2018; Cole et al. 2016; Conway et al. 2017); while other measures drawing attention to the avatar, such as the appearance of the avatar before the dots, may achieve the same effect (Bukowski et al. 2015; Gardner et al. 2018b). The results from Experiment 4 reported here suggest that the perception of agency of the avatar may be an alternative method of drawing attention to the avatar. They also provide further evidence against the automaticity of either a perspective-taking or directional orienting effect in the DPT, but combined with the results of an explicit task found in Experiment 1, suggest that ongoing attention drawn to the avatar leads to a rapid, involuntary (spontaneous) perspective-taking effect.

The simplified scene design used in Experiment 4 makes it impossible to determine whether the altercentric effect we observe here represents perspective-taking or directional orienting. This distinction requires an occlusion task, and these results suggest that the implicit occlusion tasks in Experiments 1 and 2 may have produced null results because of the use of Lego stimuli. We therefore conducted an implicit occlusion task using the original DPT avatars, to establish whether the effect found in Experiment 4 is best explained by directional orienting or perspective-taking; and whether the null results in the implicit tasks of Experiments 1 and 2 can be attributed to the stimuli used.

Experiment 5: occlusion task with original stimuli

Materials and methods

The original DPT stimuli were edited to create barriers in the same positions as in the Lego stimuli (see Figure 12). Because the new scene layout required dots to be displayed in positions other than on a flat wall, floating red orbs were used instead of the red discs used in the original DPT. A colour picking tool was used to create the

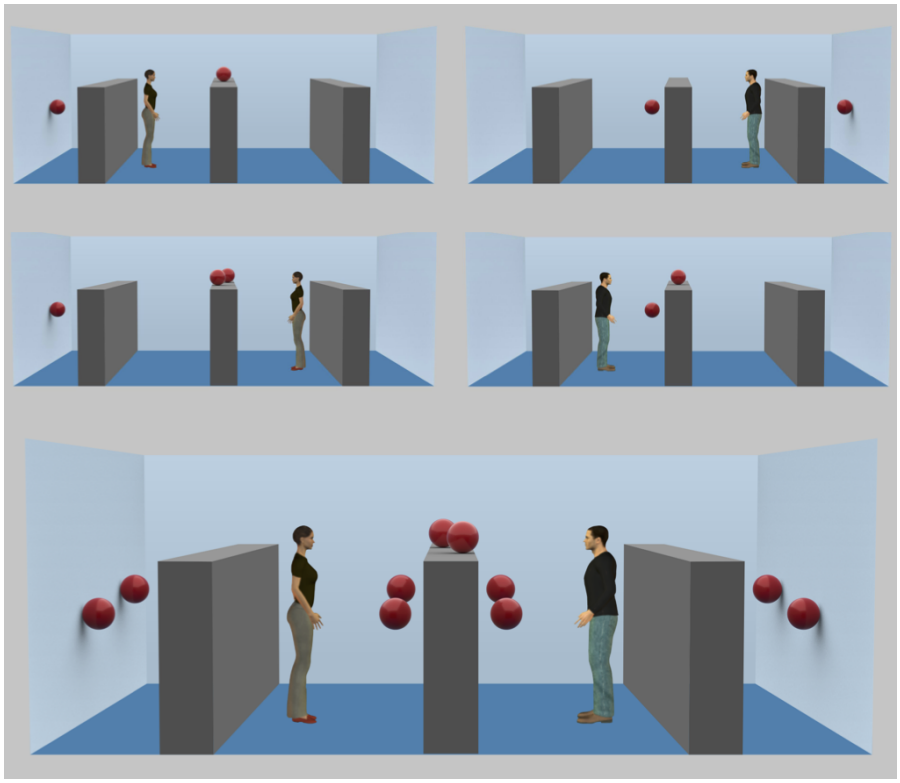


Figure 12. Occlusion task using avatars from the original DPT. The upper four images show example scenes; note that each scene that participants saw featured a single avatar and a maximum of four balls. The lower image shows both potential placement positions for avatars (left or right of the central table) and all possible ball positions (5 possible positions, maximum of two balls in any one position, and maximum of four balls per scene).

colour for the orbs, and shadows were added to create depth. They were positioned within the eyeline of the avatars, in the same positions as in the Lego stimuli.

Because this task differed from Experiment 2 only in images used, and in the position of the fixation crosses and pre-trial instructions, it was not preregistered separately, as all other details of the preregistration were the same.

Participants. We used the same sample size as in Experiments 1 to 4. Thirty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 30 minutes. Data were excluded from one participant whose task was interrupted by disconnection of the response box, and one participant who had participated in an earlier DPT. Data from these two participants was replaced by new participants. Participants gave written consent, including consent for anonymised data

Table 8. Results of Experiment 5

Slope	β	SE	χ^2	df	p
Behind avatar vs *Avatar sees*	0.021	0.005	17.97	1	$< .001^{***}$
Avatar faces (central) vs *Avatar sees*	−0.0007	0.005	0.03	1	.86
Behind avatar vs *Avatar faces (peripheral)*	0.003	0.005	0.4	1	.53

to be shared publicly. Ethical approval was granted by the University of Edinburgh's PPLSREC, reference number 188-1718/2.

Procedure. This task used the same procedure and task design as the implicit condition in Experiment 1.

Results

We applied the data exclusions and transformations described in Experiment 1, removing timed-out trials (0.40%, $n = 31$) and erroneous responses (2.51%, $n = 192$). There were no responses below 100 ms.

Following the analysis used in Experiments 1 and 2, data limited to the three relevant comparisons (*Behind avatar* vs *Avatar sees*, *Avatar faces (central)* vs *Avatar sees*, and *Avatar faces (peripheral)* vs *Behind avatar*) were modelled using three models identical to those used in Experiment 2.

As in Experiments 1 and 2, and as predicted, the models showed a significant difference between *Avatar sees* and *Behind avatar* (23.88 ms), and no significant difference between *Avatar faces (central)* and *Avatar sees* trials (−0.76 ms, see Table 8). However, contrary to our predictions, there was also no significant difference between *Avatar faces (peripheral)* and *Behind avatar* trials, at 3.31 ms (Figure 13). This suggests that there was no directional orienting effect in this task, and that the difference between *Avatar sees* and *Behind avatar* trials was driven by the spatial distribution of the balls.

Discussion

These results do not support the hypothesis that directional orienting played any role in this implicit task. This continues to conflict with findings of altercentric effects in implicit tasks (Samson et al. 2010; Santiesteban et al. 2014; Langton 2018; Cole et al. 2016; Conway et al. 2017), and is difficult to explain. One important difference between Experiment 5 and other implicit tasks is the visual complexity of the scene: where other implicit tasks have used an avatar in a consistent position within the scene, we have used two avatars in two possible positions; and where previous tasks have used goggles, a telescope, a single barrier, or a pair of barriers (one behind and one in front of the avatar), we have three different barriers in our scene, two of which are in front of the avatar. The addition of the second barrier, and the distance between the avatar and any balls behind this barrier on the periphery of the scene, may be sufficient to prevent directional orienting. The visual complexity of this scene design may therefore simply

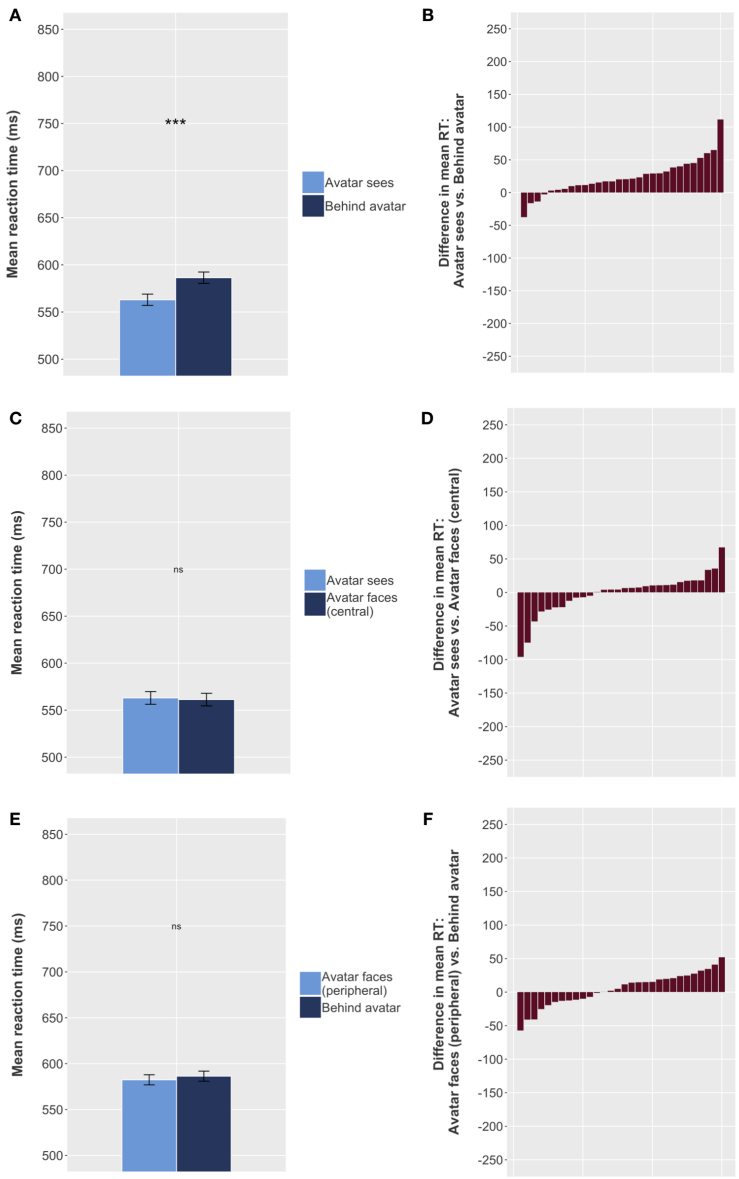


Figure 13. Results of Experiment 5. (A) Mean RT for *Behind avatar* and *Avatar sees* conditions; error bars indicate 95% CIs on the mean of the by-participant means. (B) Each individual participant's difference between mean *Behind avatar* RT and mean *Avatar sees* RT. Mean reaction time is higher (i.e. participants respond more slowly) for *Behind avatar* trials (A); a majority of participants show this effect (B). However, there is no difference in RT between *Avatar sees* and *Avatar faces (central)* (C, D) or *Behind avatar* and *Avatar faces (peripheral)* (E, F).

be too high for directional orienting to play a role in participants' comprehension of each image.

It would be instructive to replicate Experiment 1 (explicit, implicit and uncued conditions) using a range of different stimuli, including a simple screen with a window that may be open or closed (Cole et al. 2016), a scene with more realistic depth in the third dimension but still only one barrier (Baker et al. 2016), pictures of humans or human experimenters (Langton 2018), and alternative occlusion methods such as opaque goggles (Conway et al. 2017; Furlanetto et al. 2016). Further manipulations such as colour saturation and the spacing of dots may also be useful. It is clear that properties of the stimuli affect results in the DPT in a variety of ways, and exploring these effects systematically would greatly clarify the nature and triggering conditions of the altercentric effect. Given the clear range of individual participant differences in responses to the tasks, it may also be the case that much of the DPT literature is underpowered and suffers from sampling error; further research into the individual differences underlying participant responses would be valuable.

The results of Experiment 5 yield no further evidence on whether perspective-taking or directional orienting underlies the altercentric effect found in Experiment 4. It may be the case that the visual complexity of this occlusion task is too high to induce an effect on an implicit task, and that this paradigm is therefore not able to determine whether a consistency effect on a simple implicit task is the result of perspective-taking or directional orienting.

Conclusions & discussion

The results of these five experiments collectively provide evidence that differences in stimuli and task demands, and particularly in perception of the agency and relevance of the on-screen characters, play a substantial role in mediating the results of the DPT. That is, when avatars are more humanoid and realistic, they are more likely to create an altercentric effect, but only in a task of sufficient visual simplicity; and when the avatar's perspective is relevant to the task, it drives a perspective-taking effect.

Experiment 1 showed that uncued tasks (as predicted) do not result in altercentric interference; and that explicit versions of the DPT (Samson et al. 2010; Baker et al. 2016; Capozzi et al. 2014; Furlanetto et al. 2016; Marshall et al. 2018; Mattan et al. 2015, 2016; Wilson et al. 2017) likely do provide evidence of visual perspective-taking, rather than directional orienting. This coheres with our analysis of the literature as containing discrepant findings based on varying implementations of the DPT; namely, that explicit tasks find results consistent with visual perspective-taking rather than directional orienting.

This "visual perspective-taking" could plausibly be achieved by different mechanisms – for instance, by participants spatially representing the dots/discs that are visible from a certain point in the room, regardless of what occupies this position; or by representing the visual perspective of an on-screen figure. The use of a control condition using non-social stimuli such as arrows, lamps or cameras in an explicit occlusion task could be useful in distinguishing between these mechanisms. That is, if there is a perspective-taking effect on an explicit task for humanoid stimuli, but

not for non-social stimuli, it would suggest that the effect is driven by perspective-taking specific to stimuli that represent a human-like perspective. If, however, a perspective-taking effect is found regardless of stimulus type, this would suggest a spatial representation effect. It is important to note, though, that on-screen avatars have no perspective to represent (they are avatars, not agents), and so perhaps it should be expected that avatars and non-social stimuli would show similar results. It is also possible that spatial representation may be the primary mechanism by which visual perspective-taking is achieved. This would be a fruitful avenue for further research.

This visual perspective-taking is not purely stimulus-driven, instead requiring that participants are motivated to take the perspective of the avatars throughout the task. Given this continuous perspective-taking, it seems that participants maintain awareness of the avatar's perspective (which is relevant on some scenes) throughout the experiment (even on scenes where it is not relevant), and therefore use the avatar's perspective as a cue throughout the task. Mean RT for the explicit condition (720.91 ms) was higher than implicit (594.37 ms) or uncued (578.00 ms) conditions; the experiment was not powered to determine whether this between-subjects difference was statistically significant, but confirmatory research analysing this would be informative, as slower responses on an explicit task could indicate that holding the avatar's perspective in working memory is somewhat effortful. The evidence from Experiment 1 suggests that visual perspective-taking should not be considered automatic, but rather spontaneous, occurring only when relevant; but may still occur involuntarily and rapidly, on trials where it is not necessary for the immediate task (recall that all of our analyses are conducted on trials where participants are only required to take their own perspective).

Although we predicted that the implicit task in Experiment 1 would show directional orienting effects, our results in Experiments 1–3 and 5 failed to match previous findings of directional orienting in implicit versions of the DPT (Santiesteban et al. 2014; Cole et al. 2016; Conway et al. 2017; Langton 2018). Experiments 2 and 3 investigated whether this could be attributed to (failure to) draw attention to the avatars by placement of the fixation cross and pre-trial instructions, or by the greater scene complexity of the stimuli with multiple barriers, two avatars in different positions, and up to four balls. In Experiment 2 we used the fixation cross and instructions to draw attention to the avatar in an implicit task, and still found no evidence of an altercentric effect consistent with the avatar driving directional orienting; the only effect present was better explained by the spatial distribution of the scene. Likewise, in Experiment 3 we simplified our scenes and still found no altercentric effect in an implicit task. However, in Experiment 4, an implicit task using the original DPT stimuli and otherwise identical to Experiment 3 did find an altercentric effect, suggesting an (unanticipated) sensitivity of implicit tasks to the details of the on-screen characters (that is, cartoon stimuli yield interference, Lego characters do not).

Because of the simplified stimuli, it is not possible to determine whether the altercentric effect found in Experiment 4 was a result of perspective-taking or directional orienting. We therefore conducted an occlusion task using the original avatar stimuli, and otherwise identical to the implicit task in Experiment 1. This task found no evidence of directional orienting (or perspective-taking), with the only effect best explained by the spatial distribution of the scene. This battery of experiments

therefore did not confirm one of our main predictions, which was that implicit occlusion tasks would produce an altercentric effect consistent with directional orienting. The greatly increased visual complexity of Experiment 5 stimuli compared to previous implicit occlusion tasks (Cole et al. 2016; Conway et al. 2017; Langton 2018) may explain why we did not find a directional orienting effect. The unexpected results for the battery of implicit tasks presented in this paper suggest the need for future research exploring the variety of ways in which DPT stimuli may affect the results, and the theoretical implications of these variations.

Collectively, these five experiments point to a complex picture of visual perspective-taking, as something occurring spontaneously in dynamic reaction to the immediate environment, based on attentional cues. Our Experiment 1 provides evidence that explicit versions of the DPT likely do provide evidence of visual perspective-taking, rather than directional orienting. The contrast between explicit and implicit/uncued conditions suggests that visual perspective-taking is not purely stimulus-driven, instead requiring that participants are motivated to take the perspective of the avatars throughout the task. Visual perspective-taking should therefore not be considered automatic, but rather spontaneous, occurring only when relevant.

Our results across all five experiments also suggest that the visual complexity of the scene and the perceived agency of the stimuli play a role in driving the appearance of an altercentric effect, contributing further evidence that directional orienting is not automatic, and is instead potentially dependent on the directional cues provided by the stimulus, or on cues to consider the agent's perspective as relevant (albeit not sufficiently to sustain throughout the task, as in an explicit task). Given this result we emphasise that a clear distinction should be made between processes that are automatic and processes that are spontaneous – that is, not automatic but still rapid, involuntary, and unconscious, arising when necessary, as prompted by the attentional system.

The possibility that perspective taking might be spontaneous raises an important theoretical issue. Specifically, it raises the possibility that directional orienting and perspective taking are in fact not cognitively distinct alternatives. Instead there might be more of a continuum between them, by which directional orienting is a possible input to perspective-taking, with its effect modulated by attention. This possibility is an important topic for future research, both theoretical and empirical. Finally, we note that while the DPT is proving a fruitful method for exploring questions regarding visual perspective-taking, our results suggest that caution is required to interpret results from a range of tasks with widely varying stimuli and implementation. Given the application of this task to broader questions about theory of mind (Drayton et al. 2018; Yue et al. 2017), it is essential to clarify the precise causes of altercentric interference before using this task to establish group differences in, or the presence or absence of, perspective-taking.

Acknowledgements

We would like to thank our colleagues Simon Kirby and Jennifer Culbertson, who provided insight and expertise that greatly assisted this research; and Rachel Kindellan, for her assistance in collecting data for Experiment 5.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681942. TSP was financially supported by Durham University's Addison Wheeler bequest and by the European Research Council, under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 609819 (Somics project)). COG was financially supported by the University of Edinburgh's Principal's Career Development Scholarship, Global Research Scholarship, and Research Support Grants.

References

Apperly I (2011) *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Hove [East Sussex] ; New York: Psychology Press. ISBN 978-1-84169-697-3. OCLC: ocn432998292.

Baayen H and Milin P (2010) Analyzing reaction times. *International Journal of Psychological Research* 3(2): 12. DOI:10.21500/20112084.807.

Baker LJ, Levin DT and Saylor MM (2016) The extent of default visual perspective taking in complex layouts. *Journal of Experimental Psychology: Human Perception and Performance* 42(4): 508–516. DOI:10.1037/xhp0000164.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* 67(1). DOI:10.18637/jss.v067.i01.

Bukowski H, Hietanen JK and Samson D (2015) From gaze cueing to perspective taking: Revisiting the claim that we automatically compute where or what other people are looking at. *Visual Cognition* 23(8): 1020–1042. DOI:10.1080/13506285.2015.1132804.

Capozzi F, Cavallo A, Furlanetto T and Becchio C (2014) Altercentric Intrusions from Multiple Perspectives: Beyond Dyads. *PLoS ONE* 9(12): e114210. DOI:10.1371/journal.pone.0114210.

Carruthers P (2017) Mindreading in adults: Evaluating two-systems views. *Synthese* 194(3): 673–688. DOI:10.1007/s11229-015-0792-3.

Cole G, Atkinson M, D'Souza A and Smith D (2017) Spontaneous Perspective Taking in Humans? *Vision* 1(2): 17. DOI:10.3390/vision1020017.

Cole G, Atkinson M, Le AT and Smith DT (2016) Do humans spontaneously take the perspective of others? *Acta Psychologica* 164: 165–168. DOI:10.1016/j.actpsy.2016.01.007.

Conway JR, Lee D, Ojaghi M, Catmur C and Bird G (2017) Submentalizing or mentalizing in a Level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance* 43(3): 454–465. DOI:10.1037/xhp0000319.

Drayton LA, Santos LR and Baskin-Sommers A (2018) Psychopaths fail to automatically take the perspective of others. *Proceedings of the National Academy of Sciences* 115(13): 3302–3307. DOI:10.1073/pnas.1721903115.

Freundlieb M, Kovács AM and Sebanz N (2016) When do humans spontaneously adopt another's visuospatial perspective? *Journal of Experimental Psychology: Human Perception and Performance* 42(3): 401–412. DOI:10.1037/xhp0000153.

Freundlieb M, Kovács AM and Sebanz N (2018) Reading Your Mind While You Are Reading—Evidence for Spontaneous Visuospatial Perspective Taking During a

Semantic Categorization Task. *Psychological Science* 29(4): 614–622. DOI:10.1177/0956797617740973.

Furlanetto T, Becchio C, Samson D and Apperly I (2016) Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance* 42(2): 158–163. DOI:10.1037/xhp0000138.

Gardner M, Bileviciute A and Edmonds C (2018a) Implicit Mentalising during Level-1 Visual Perspective-Taking Indicated by Dissociation with Attention Orienting. *Vision* 2(1): 3. DOI:10.3390/vision2010003.

Gardner MR, Hull Z, Taylor D and Edmonds CJ (2018b) 'Spontaneous' visual perspective-taking mediated by attention orienting that is voluntary and not reflexive. *Quarterly Journal of Experimental Psychology* 71(4): 1020–1029. DOI:10.1080/17470218.2017.1307868.

Heyes C (2014) Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science* 9(2): 131–143. DOI:10.1177/1745691613518076.

Langton S (2018) I Don't See It Your Way: The Dot Perspective Task Does Not Gauge Spontaneous Perspective Taking. *Vision* 2(1): 6. DOI:10.3390/vision2010006.

LaPointe MRP, Cullen R, Baltaretu B, Campos M, Michalski N, Sri Satgunarajah S, Cadieux ML, Pachai MV and Shore DI (2016) An attentional bias for LEGO® people using a change detection task: Are LEGO® people animate? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 70(3): 219–231. DOI:10.1037/cep0000077.

Marotta A, Lupiáñez J, Martella D and Casagrande M (2012) Eye gaze versus arrows as spatial cues: Two qualitatively different modes of attentional selection. *Journal of Experimental Psychology: Human Perception and Performance* 38(2): 326–335. DOI:10.1037/a0023959.

Marshall J, Gollwitzer A and Santos LR (2018) Does altercentric interference rely on mentalizing?: Results from two level-1 perspective-taking tasks. *PLOS ONE* 13(3): e0194101. DOI:10.1371/journal.pone.0194101.

Mattan B, Quinn KA, Apperly IA, Sui J and Rotshtein P (2015) Is it always me first? Effects of self-tagging on third-person perspective-taking. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(4): 1100–1117. DOI:10.1037/xlm0000078.

Mattan BD, Rotshtein P and Quinn KA (2016) Empathy and visual perspective-taking performance. *Cognitive Neuroscience* 7(1-4): 170–181. DOI:10.1080/17588928.2015.1085372.

Michael J, Wolf T, Letesson C, Butterfill S, Skewes J and Hohwy J (2018) Seeing it both ways: Using a double-cuing task to investigate the role of spatial cuing in Level-1 visual perspective-taking. *Journal of Experimental Psychology: Human Perception and Performance* 44(5): 693–702. DOI:10.1037/xhp0000486.

Moors A and De Houwer J (2006) Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin* 132(2): 297–326. DOI:10.1037/0033-2909.132.2.297.

Nielsen MK, Slade L, Levy JP and Holmes A (2015) Inclined to see it your way: Do altercentric intrusion effects in visual perspective taking reflect an intrinsically social process? *Quarterly Journal of Experimental Psychology* 68(10): 1931–1951. DOI:10.1080/17470218.2015.1023206.

Peirce J (2010) PsychoPy - Psychology software for Python : 289.

Qureshi AW, Apperly IA and Samson D (2010) Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition* 117(2): 230–236. DOI:10.1016/j.cognition.2010.08.003.

{R Core Team} (2015) R: A language and environment for statistical computing .

Samson D, Apperly IA, Braithwaite JJ, Andrews BJ and Bodley Scott SE (2010) Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance* 36(5): 1255–1266. DOI: 10.1037/a0018729.

San Juan V and Astington JW (2017) Does language matter for implicit theory of mind? The effects of epistemic verb training on implicit and explicit false-belief understanding. *Cognitive Development* 41: 19–32. DOI:10.1016/j.cogdev.2016.12.003.

Santiesteban I, Catmur C, Hopkins SC, Bird G and Heyes C (2014) Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance* 40(3): 929–937. DOI:10.1037/a0035175.

Simmons JP, Nelson LD and Simonsohn U (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22(11): 1359–1366. DOI:10.1177/0956797611417632.

Singmann H, Bolker B, Westfall J and Aust F (2017) Afex: Analysis of Factorial Experiments. *R package version 0.17-8.* .

Surtees A, Apperly I and Samson D (2016) I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition* 150: 43–52. DOI:10.1016/j.cognition.2016.01.014.

Surtees ADR and Apperly IA (2012) Egocentrism and Automatic Perspective Taking in Children and Adults: Egocentrism and Automatic Perspective Taking. *Child Development* : no–noDOI:10.1111/j.1467-8624.2011.01730.x.

Trick L and Pylyshyn Z (1994) Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review* .

Westra E (2017) Spontaneous mindreading: A problem for the two-systems account. *Synthese* 194(11): 4559–4581. DOI:10.1007/s11229-016-1159-0.

Whelan R (2008) Effective Analysis of Reaction Time Data. *The Psychological Record* 58(3): 475–482. DOI:10.1007/BF03395630.

Wilson CJ, Soranzo A and Bertamini M (2017) Attentional interference is modulated by salience not sentience. *Acta Psychologica* 178: 56–65. DOI:10.1016/j.actpsy.2017.05.010.

Yue T, Jiang Y, Yue C and Huang X (2017) Differential Effects of Oxytocin on Visual Perspective Taking for Men and Women. *Frontiers in Behavioral Neuroscience* 11. DOI: 10.3389/fnbeh.2017.00228.

Zhao X, Cusimano C and Malle BF (2015) In Search of Triggering Conditions for Spontaneous Visual Perspective Taking. *CogSci* : 6.

5.2 Supplementary methods

The Supplementary Information included with O'Grady et al. (submitted) included a brief presentation of the arrow/avatar DPT (presented in more detail in this thesis in Chapter 4), and further details on the methods of Experiments 1 and 3. These sections of the SI are included below, before an extended discussion on the results of Experiments 1 to 5.

5.2.1 Detailed Experiment 1 methods

The full range of constraints followed in balancing Experiment 1 was as follows. Sally and Andrew each appear in half of the trials, randomly on either the left-hand side or right-hand side of the screen; and the sides of Yes and No were counterbalanced across participants, but remained consistent for a given participant. In the explicit condition only, there were 128 Self trials and 128 Other trials per participant. The implicit and uncued conditions each had 256 Self trials, to ensure that the tasks were the same length.

Half of all trials required a Yes response from participants; half required a No response. In line-of-sight consistent trials, any digit that does not match the number of balls seen by both the participant and avatar should generate a No response. In line-of-sight inconsistent trials, a No digit might match the perspective not being assessed in that trial (e.g. the avatar's perspective on a Self trial), or a number of balls seen by neither the participant nor the avatar. In line-of-sight inconsistent trials, No trials were therefore further divided into half No–other and half No–none (in the case of Self trials), and half No–self and half No–none (in the case of Other trials) (see Figure 5.1 and Figure 5.2).

Each possible number of balls (1, 2, 3 or 4) was presented 64 times. Each avatar, consistency, self vs other condition, and yes vs no condition was balanced across the number of balls; for example, there is one No–other self *Avatar sees* trial with Sally and one ball present, and one No–none trial in an otherwise identical combination of conditions. In addition, there were 16 filler trials per avatar with zero balls. In some cases, there were multiple possible scenes that fulfilled a particular combination of conditions; in those cases, one scene was randomly selected when the trial list for that participant was generated.

5.2.2 Detailed Experiment 3 methods

The full range of constraints followed in balancing Experiment 3 was as follows. Sally and Andrew each appeared in half of the trials, each facing the left-hand side of the screen on half of

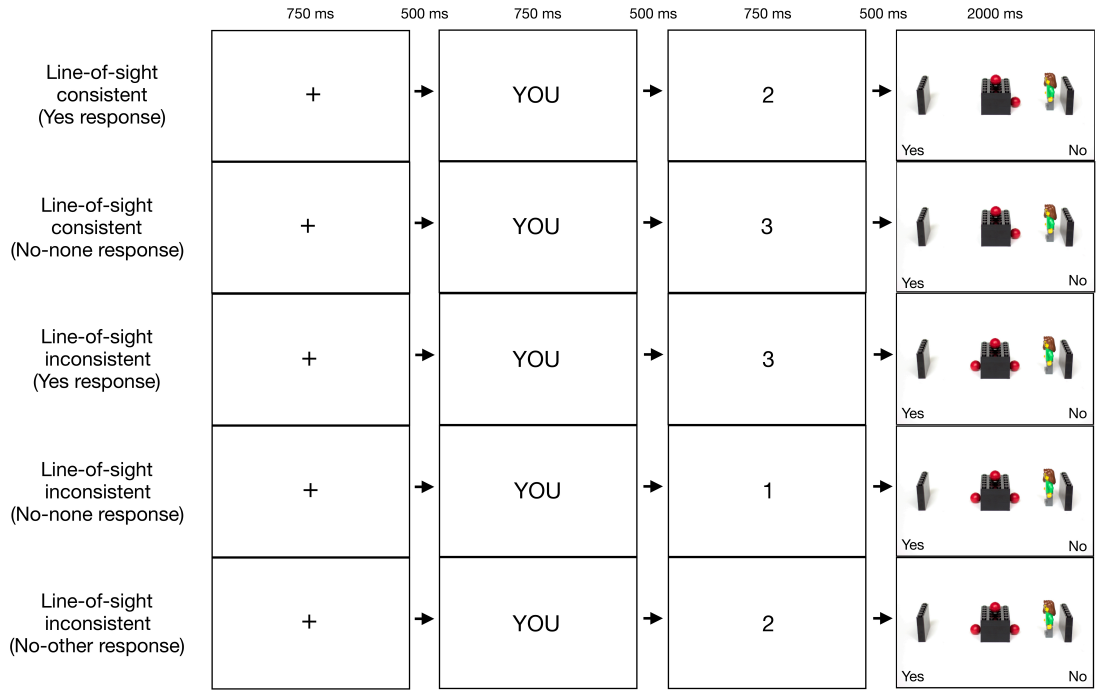


Figure 5.1: An illustration of the trial procedure for Self trials (fixation cross, word, digit, scene) with examples of Yes and No–none options in the line-of-sight consistent scenes, and Yes, No–none and No–other options in the line-of-sight inconsistent scenes (which includes both *Avatar faces* and *Behind avatar* scenes).

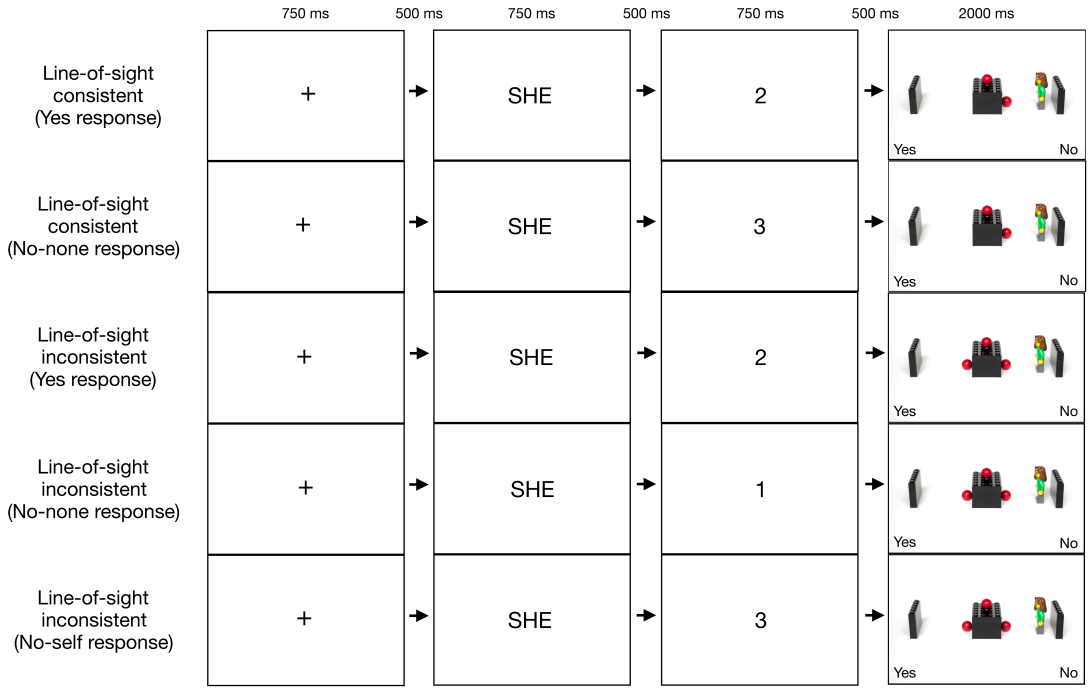


Figure 5.2: An illustration of the trial procedure for Other trials (fixation cross, word, digit, scene) with examples of Yes and No–none options in the line-of-sight consistent scenes, and Yes, No–none and No–self options in the line-of-sight inconsistent scenes.

those trials, and the right-hand side on the other half. Half of all trials required a Yes response from participants; half required a No response; in *Behind avatar* trials, half of the No response trials were No–none and half were No–other. Half of the trials had one ball, and half had two. Each avatar, consistency, and yes vs no condition was balanced across the number of balls; for example, there were six No–other *Behind avatar* trials with Sally and one ball present, and six No–none trials in an otherwise identical combination of conditions. In addition, there were six filler trials per avatar with zero balls.

5.3 Extended discussion

As discussed in Chapter 3, the limitations of rapid mindreading in adults are relevant to assessing the plausibility of the ostensive-inferential model of communication. Evidence of automatic or spontaneous perspective-taking in particularly relevant given its utility in testing the predictions of the two-systems and submentalising accounts of mindreading. This is because a two-systems account considers automatic perspective-taking to belong to System 1 – that is, informationally encapsulated, and therefore not subject to top-down reasoning, or capable of providing information available to System 2 processes. The submentalising account argues that evidence of spontaneous mindreading is better explained by behaviours that give the appearance of mindreading, but are achieved by lower-level cognitive processes like attentional orienting. Neither of these accounts would allow for the kind of rapid, flexible and context-dependent mindreading necessitated by ostensive communication.

The work presented in this chapter tested predictions made by both accounts about spontaneous/automatic perspective-taking, finding in both cases that these predictions were not supported by the findings:

1. The submentalising account argues that the altercentric effect is best explained by attentional orienting rather than perspective-taking. It therefore predicts that the avatar's direction of facing, rather than visual perspective, drives the altercentric effect. On an occlusion task, this would be demonstrated by scenes in which an avatar faces, but does not see, certain balls nonetheless being treated as "consistent" and having no difference in reaction times from scenes in which an avatar can see all the balls. In a spatially controlled task, we found no evidence that the avatar's direction of facing caused attentional orienting, but did find that the avatar's visual perspective created an altercentric effect. This suggests that the behaviour identified in the DPT is a genuine calculation of the avatar's perspec-

tive, and cannot be explained by submentalising. We found little evidence of directional orienting altogether; only the results of Experiment 4 showed an effect on an implicit task, which could be consistent with either directional orienting or perspective-taking; and a null result on Experiment 5 failed to distinguish between these explanations.

2. The two-systems account argues that System 1 mindreading processes are characterised by, among other things, informational encapsulation. This account suggests that the altercentric effect in the DPT, if it is truly an encapsulated System 1 process, should not depend on access to high-level goals, expectations and beliefs, such as beliefs about the avatar's perspective and whether it is relevant to the task. While encapsulated processes may have access to certain kinds of top-down information flow from related processes (such as the flow from auditory processing and lexical knowledge to phonological identification), general background knowledge and beliefs are not accessible to encapsulated processes on the canonical definition of informational encapsulation (Fodor 1983; Coltheart 1999). Beliefs about the avatar and its relevance to the task should therefore, on this account, not either cause or disrupt the altercentric effect. However, we find that the requirement to take the avatar's perspective is the crucial factor that induces an altercentric effect: it is only in an explicit task that we find a perspective-taking effect, suggesting that it is attention to the avatar—based on knowledge and beliefs about the task—that drives perspective-taking. Features of the avatar also appear to alter the findings, with an altercentric effect on an implicit task appearing only when the original stimuli were used, suggesting that beliefs about what the on-screen avatar is intended to represent may play a role in inducing an effect. Neither of these results is consistent with the informational encapsulation of the two-systems account.

Westra (2017b, p. 11) argues that evidence of spontaneous mindreading that is sensitive to attention creates a problem for the two-systems account precisely because of this claim of informational encapsulation:

> *"By acknowledging a role for attention in perspective-selection, two-systems theorists are opening up a space where goals might play a significant role in the Level-1 perspective-taking system."*

The results presented in this chapter contribute to a growing literature that shows that spontaneous mindreading (including both Level 1 and Level 2 perspective-taking) is sensitive to goal-directed behaviour, including attention and background knowledge.

In the gaze-cueing paradigm, a robot face has been found to be less likely to induce a congruence effect than a human face, unless participants are told that a human experimenter is controlling the robot's face; in this case, the robot face is as likely to induce a congruence effect as the human face (Wiese et al. 2012). When the face cue is ambiguous – plausibly representing either wheels on a car or human eyes – the congruence effect depends on whether participants were told that the face cue represented wheels or eyes (Ristic and Kingstone 2005).

Other paradigms have found that participants' beliefs about the other agent's goals, intentions and agency play a role in inducing perspective-taking. Elekes et al. (2016) found evidence of spontaneous perspective-taking on a version of the Level 2 DPT, but only under certain conditions. In this modified task, participants completed a number task alone or with another participant. The task involved judging whether a number on a table in front of them (0 or 8 for the consistent condition; 6 or 9 for the inconsistent condition) matched the number read in a recording. In the "joint" condition, the partner either completed the same task, or a different task that required judging the colours of the numbers. That is, only certain participants in the "joint" condition believed that their partner was paying attention to the value of the number rather than its colour. The task did not require any actual cooperation, making any perspective-taking spontaneous. Only participants whose partners were engaged in the same task showed interference from their partner's perspective. In a very similar task, Surtees et al. (2016a) found spontaneous Level 2 perspective-taking with a live partner. These results contribute to the evidence from our Experiment 3 and Experiment 4 that the kind of agent used in a perspective-taking task makes a difference to the results.

In both of these studies, simply having another agent with a perspective on the numbers on the table was not sufficient to trigger perspective-taking: it was only when participants considered their partner's perspective to be relevant to the situation, even if not to their own immediate task, that they spontaneously considered that perspective throughout the task. These results also contradict the original DPT findings that Level 1, but not Level 2, perspective-taking occurs spontaneously, further challenging the "informational encapsulation" premise of the two-systems account of the difference between Level 1 and Level 2.

Different paradigms provide further evidence that spontaneous perspective-taking may be induced by top-down processes. Zwickel (2009), as discussed in Chapter 3, find that participants spontaneously take the perspective of a triangle only after watching an animation in which the triangle had behaved as if it had mental states, but not an animation in which the triangle moved aimlessly. Freundlieb et al. (2016) find that participants adopt the visual perspective of a con-

federate seated at a 90° angle to them when completing a reaction time task, but only if they perceived the confederate as acting intentionally; and in a related task, only if the confederate's visual access to the scene was not hampered by blinding glasses (Freundlieb et al. 2017). A similar task found that this spontaneous perspective-taking effect occurred even when the task was word-reading, involving a complex learned skill and visually complex stimulus (Freundlieb et al. 2018).

In a similar vein, a range of more recent DPT results contribute to the growing picture that Level 1 perspective-taking is not automatic, but depends on a range of top-down processes including attention and even group membership (Ferguson et al. 2018; Gardner et al. 2018b; Schneider et al. 2018). This growing body of evidence suggests that both Level 1 and Level 2 perspective-taking can be spontaneous, and induced by a range of factors, including the characteristics of the agent and attention directed to the agent. This evidence challenges the two-systems account.

5.4 Chapter summary

This chapter presented a series of five experiments showing that the altercentric effect is not automatic; that is, not purely stimulus-driven. By using barriers that concealed balls from avatars' line of sight, these tasks compared the directional orienting vs perspective-taking explanations of the altercentric effect. An experiment comparing explicit, implicit, and uncued variants of the task found an altercentric effect in the explicit condition that was best accounted for by perspective-taking rather than directional orienting. There was no directional orienting or perspective-taking effect on either implicit or uncued tasks.

Given the unexpected null result in the implicit condition, a range of follow-up experiments attempted to establish the reason for this finding. Following evidence that drawing attention to the avatar on a trial-by-trial basis has prompted an altercentric effect even on an uncued task, a second implicit task moved the pre-trial cues to appear over the point at which the avatar would appear in the scene, but found no altercentric effect. A considerably simplified implicit task, reducing the scene to just a Lego character and up to two balls, similarly found no effect, suggesting that the visual complexity of the scene could not explain the null result. An exact replication of this simplified implicit Lego task using the avatars from the original DPT did find an effect, suggesting that some feature of the avatars (whether perception of agency, directional properties, or other features) plays a role in prompting an effect on implicit tasks. The effect

on this task could be consistent with either a perspective-taking or directional account, which prompted a replication of the implicit condition of Experiment 1 (an implicit occlusion task with the original complex scene layout) using the original avatars. This task found no altercentric effect, suggesting limits on the complexity of implicit tasks that are capable of prompting an altercentric effect.

These five tasks found no evidence for an automatic perspective-taking or directional orienting effect. The finding of a perspective-taking effect in an explicit task suggests that perspective-taking is spontaneous – that is, rapid and involuntary – but not automatic, since it is reliant on sufficient motivation to take the avatar's perspective throughout the task. Collectively, the results presented in this chapter challenge both the submentalising account (which suggests that visual perspective-taking in the DPT is achieved by submentalising behaviour such as directional orienting); and the two-systems account (which suggests that visual perspective-taking should be automatic and informationally encapsulated). The following chapter explores the altercentric effect underlying these results, and possible explanations for it, in greater detail.

Chapter 6

The altercentric effect: processing costs or preferential attention?

The evidence presented in the previous two chapters suggests that spontaneous perspective-taking is context-dependent, occurring when attention, task demands, and the features of other agents require it. This provides evidence of rapid, involuntary perspective-taking. The presence of the altercentric effect nonetheless suggests that there is a cost – albeit a small one – to processing a conflicting visual perspective, as opposed to a shared visual perspective. The cause of this cost is poorly understood, with little research having investigated the nature of the delay itself, rather than which experimental designs induce the delay.

Given the ubiquity of mindreading in human social interaction, evidence of inefficiency or effort in mindreading is somewhat unexpected. As discussed in Chapter 3, it is this contradiction that different accounts of mindreading attempt to reconcile. The submentalising account does so by arguing that the *appearance* of mindreading does not imply *actual* mindreading, and that many of the social interactions that appear to require mindreading may in fact rely on submentalising instead. The two-systems account argues that a limited subset of mindreading abilities is efficient, inflexible and automatic, while the remainder is flexible and powerful, but less efficient. The one-system, mentalising account upon which the ostensive-inferential model rests suggests instead that mindreading may appear to be inefficient and effortful, but that this understanding is mistaken and there is a better explanation for evidence suggesting inefficient or effortful mindreading.

The altercentric delay suggests that representing another agent's visual perspective incurs a small but consistent processing cost. This is not fatal to either the two-systems account or the ostensive-inferential account, but does pose the question of how it can be explained by these

accounts. It does not pose the same question for the submentalising account, which would expect a processing cost for mindreading behaviour.

Qureshi et al. (2010) suggest that the delay arises from perspective selection, rather than perspective calculation, and that this is indicative of a limit to the capacity of System 1 mindreading. That is, System 1 may efficiently calculate the perspective of another agent, but the DPT requires participants to select either their own perspective or the avatar's perspective on each trial, and it is this selection (which is beyond the capacity of System 1) that induces the delay. Therefore, on the two-systems account, the altercentric effect does not arise from perspective calculation, but nonetheless does arise from a processing cost.

I suggest an alternative explanation that is coherent with the ostensive-inferential account: that the altercentric effect arises not from a delay in the processing of conflicting perspectives, but rather because the presence of another agent drives preferential attention to the stimuli that are in a joint visual field, resulting in a delay in accounting for stimuli that are not in this field. That is, the delay may be explained not by a processing cost in reconciling conflicting perspectives, but by the direction of social attention to joint (and therefore more salient) stimuli, with a resulting decrease in salience and (and therefore reduction in processing efficiency) of stimuli that are not highlighted by this social attention.

The *processing cost* and *preferential attention* accounts of the altercentric delay predict different results from the addition of a second avatar to the DPT. The processing cost account would predict an increased delay with the addition of a second avatar, since this represents an additional perspective to process and therefore an additional source of delay. The preferential attention account would predict that a second avatar would eliminate the delay, since social attention would now be equally directed towards all stimuli, rather than preferentially directed by the presence of a single avatar with a limited perspective on the scene.

This chapter presents an extension of the task described in the previous two chapters, using two avatars simultaneously, designed to test these different accounts of the cause of the delayed response in altercentric effects.

6.1 Multiple avatars in the DPT

The addition of a second avatar runs the risk of overwhelming spontaneous perspective-taking. It may be useful to efficiently track the perspective of another individual if their perspective seems somehow relevant; it is possibly less useful, or even feasible, to simultaneously track the

perspectives of multiple individuals. For this reason, there is a chance that the altercentric effect might not be found on perspective-taking tasks that involve more than one agent. DPT variants involving more than one avatar have not been plentiful, but Capozzi et al. (2014) found that a dual-avatar task did not eliminate the altercentric effect on single-avatar trials, suggesting that spontaneous perspective-taking for single avatars should not be overwhelmed by a more complex task.

In this explicit task (Capozzi et al. 2014), either one or two avatars could appear in each scene. On single-avatar scenes, the avatar appeared either in the standard DPT position (in the middle of the room facing a wall), or angled towards the front or back corner of the room. On double-avatar scenes, the two avatars either both faced into the centre of the room, or faced the front and back corners (see Figure 6.1). Unlike the standard DPT, discs could be displayed towards the front or back corners of the room, such that an avatar facing the front corner would be able to see a disc in that corner, but not in the back corner of the same wall. In this way, two avatars could both have their backs turned to dots displayed on the wall behind them, but still each have a unique perspective of the dots arrayed on the corners of the room.

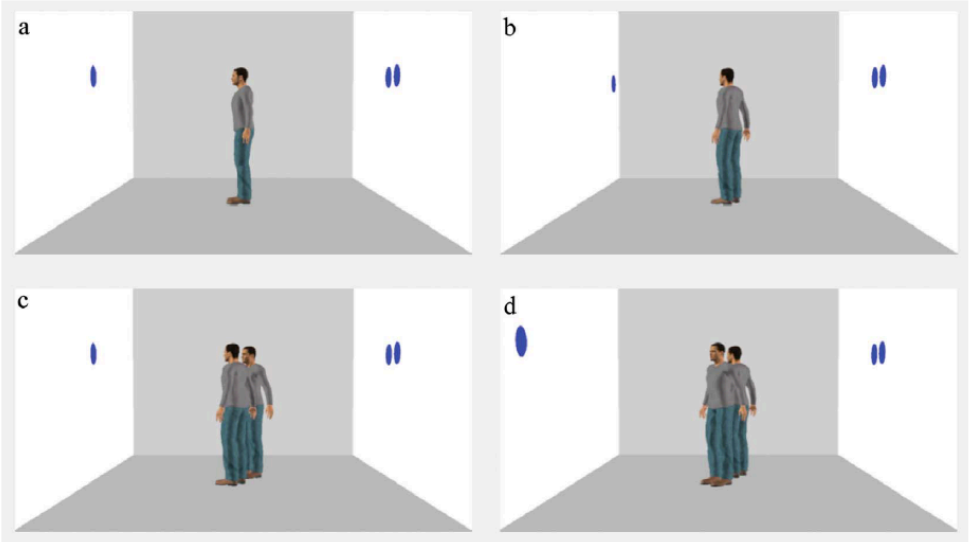


Figure 6.1: Stimuli used in Capozzi et al. (2014).

The task found the standard altercentric effect on trials with one avatar, and on trials with two avatars with a shared perspective. There was no delay when each of the two avatars had a unique perspective. This coheres with the predictions of the preferential attention account: two avatars with a shared perspective should have the same effect on preferential attention as one avatar with that perspective (i.e. driving attention to the stimuli in a joint visual field, which are the same for both these one-avatar and two-avatar scenes), but two avatars with unique

perspectives would mean no single joint visual field, no preferential attention to any particular set of dots, and therefore no delay. The results do not support the processing cost account, which would predict an increased delay with additional perspectives to process.

However, this support is tentative because the stimuli used in this experiment are arguably ambiguous enough that the perspectives of the avatars in dual-avatar scenes could easily be misconstrued: an avatar facing towards the back corner could perhaps see a dot in the front corner with its peripheral vision. More importantly, the result may have stemmed directly from the task instructions: on scenes with two avatars, each with a different perspective, participants were told to respond based on the total number of dots seen by both avatars. That is, on an "Other" perspective trial, if one avatar saw two dots and the other saw one dot, the correct response was "three". This, in addition to the fact that the two avatars were identical, may have induced participants to ignore the individual perspectives of the dual-avatar scenes in which the avatars faced different directions.

Michael et al. (2018) find that the presence of a second avatar in a DPT-like task appears to increase processing speed in some contexts, but reduce it in others. This study focused on the processing efficiency of inconsistent scenes as opposed to the difference between consistent and inconsistent scenes, and therefore did not calculate the altercentric effect. Its results are nonetheless pertinent since they suggest that the presence of a second avatar induces substantially different processing effects.

The three experiments in Michael et al. (2018) were designed to investigate the mentalising (perspective-taking) vs submentalising (referred to as spatial cueing) explanations for the altercentric effect. Based on evidence from the gaze cueing paradigm that two simultaneous cues directing attention to both sides of the screen result in faster target detection on either side of the screen than no gaze cue at all (Maylor 1985; Posner and Cohen 1984), Michael et al. (2018) propose that spatial cueing *facilitates* processing of a scene when perspectives are inconsistent, while perspective-taking *delays* processing when perspectives are inconsistent. If scenes with two back-to-back avatars are processed more efficiently than scenes with a single avatar, this would therefore provide support for the spatial cueing account; whereas on the perspective-taking account, two avatars should be less efficient than one, and one should be less efficient than zero.

This series of tasks told participants to ignore the stimulus, but did not use the redundant YOU cue, meaning that they fell somewhere between the "uncued" and "implicit" categories established in Chapter 4. They did not use the altercentric effect – the difference between consis-

tent and inconsistent trials – but instead compared performance on inconsistent trials between scenes with zero, one, and two avatars. On the one-avatar trials that were analysed, perspectives were always inconsistent, with dots on both walls (in front of and behind the avatar); on the two-avatar scenes, the avatars stood back-to-back so that each was looking at one set of dots, with no dots that were hidden from both avatars. Trials with dots only on one wall were included as distractor trials to prevent participants from anticipating dots arrayed on both walls in every trial. As in the standard DPT, rather than gaze-cueing adapted versions of the DPT, participants were shown a digit for each trial and then asked to verify whether the digit matched the number of dots in the following scene.

The first experiment, which used an SOA of 800 ms, found support for the spatial cueing account, with the dual-avatar scenes resulting in the most efficient responses. The second experiment, with no SOA, found that dual-avatar scenes were less efficient than zero-avatar scenes (with no significant difference between zero- and one-avatar scenes, or one- and two-avatar scenes). A third experiment, which used a simultaneous auditory-tone-monitoring distraction task and reintroduced the SOA, similarly found that dual-avatar scenes were less efficient than zero-avatar scenes, but that there was no difference between zero- and one-avatar scenes, or one-avatar and two-avatar scenes.

Michael et al. (2018) interpret these results to suggest that both spatial cueing and perspective-taking mechanisms work somewhat concurrently. The results of Experiments 1 and 2 suggest that perspective-taking operates earlier in the processing of an image, driving decreased efficiency with an additional avatar when there is no SOA (Experiment 2); while spatial cueing operates later, if the avatars appear before the dots, giving participants time to distribute their attention accordingly (Experiment 1). The results of Experiment 3 further suggest that this spatial cueing is disrupted by a distractor task, suggesting that it is not automatic.

These results are coherent with the evidence presented in this thesis, in that spatial cueing (referred to as directional orienting in Chapters 4 and 5) appears not to be automatic. It is not clear, though, that these tasks distinguish between *spatial* cueing and perspective-taking so much as between *attentional* cueing and perspective-taking (the distinction that is the focus of this chapter).

To understand the difference, consider the occlusion tasks that have been used to investigate spatial cueing and perspective-taking. Spatial cueing refers to the directing of attention in the direction that the avatar faces, rather than to the focus of the avatar's visual perspective. This is why occlusion tasks are a suitable vehicle for investigating the distinction between spatial cueing

and perspective-taking: where spatial cueing predicts that "consistency" is defined by items that appear simply in front of the avatar, regardless of whether the avatar can "see" them, perspective-taking predicts that "consistency" is defined by *what the avatar is looking at*.

Michael et al. (2018) do not establish whether attention is specifically focused on items seen by the avatar, as opposed to merely faced by the avatar. The finding that attention is directed to both sides of the screen by the appearance of two avatars 800 ms before any dots is therefore not an indication of spatial cueing (where this term specifically implies a submentalising cue that does not involve perspective-taking), but rather that the avatars successfully direct participants' attention to the points at which the dots will appear – that is, attentional cueing, with no data to establish whether this attention is cued through submentalising or mentalising mechanisms. The fact that attentional cueing is found *later* in the processing of a scene, and is disrupted by an additional tasks, suggests that it is preceded by, and perhaps dependent on, perspective-taking. Because these tasks were implicit/uncued, did not calculate altercentric effects, and did not use occlusion, it is not possible to say that it demonstrated perspective-taking *per se* in Experiment 2; but the evidence that attentional orienting facilitates, rather than delays, scene processing is further tentative evidence in favour of the preferential attention account.

This chapter presents two new dual-avatar experiments designed to test the preferential attention account directly, in an explicit occlusion task that compares the altercentric effect in single- and double-avatar trials.

6.2 Experiment 1

The Lego DPT described in Chapters 4 and 5 was expanded to include two avatars in the scene simultaneously. These stimuli were used because, although they had resulted in null results on implicit tasks, an explicit task had found an effect; use of the same stimuli therefore avoided manipulating more methodological differences than necessary between tasks, and offered the opportunity to replicate the finding of a single-avatar altercentric effect on an explicit Lego task.

If the altercentric effect is driven by a delay in processing conflicting perspectives, then the presence of an additional avatar should increase that delay. That is, treating a consistent scene with one avatar as a baseline reaction time, an inconsistent scene with a single avatar should be slower, as a result of the processing cost imposed by an inconsistent perspective; and an inconsistent scene with two avatars (each seeing a distinct cluster of balls) should be slower still, as a result of the increase processing cost imposed by a second avatar with a second inconsistent per-

spective. On the other hand, if the altercentric effect is driven by preferential attention to balls in a joint visual field, then an additional avatar with an inconsistent perspective should not be any slower than a consistent scene with one avatar. This is because, if each avatar has a unique perspective, this scene no longer has a grouping of balls that is marked as being particularly salient by being in a joint visual field. There is therefore no preferential attention directed to any especially salient balls, and no delay in processing the stimuli that are not within a joint visual field. Note that the preferential attention hypothesis does not imply a lack of perspective-taking: in order to identify which balls are jointly visible and therefore more salient, perspective-taking is necessary.

Given the complexity of multiple perspectives within a single scene, the terms *consistent*, *Avatar sees* etc. are insufficient; one scene may have both *Avatar faces* and *Avatar sees* perspectives. This chapter will therefore shift to using the terms *shared* and *unshared*, where *shared* means that participants and all present avatars see the same number of balls, and *unshared* means that every individual (participant and either one or two avatars) has a unique perspective.

The preferential attention account and processing cost accounts of the altercentric effect make the following predictions (see Figure 6.2 for clarity):

1. *Single shared* scenes will be faster than *Single unshared* scenes, because balls outside the shared perspective will be processed more slowly (that is, a replication of the standard altercentric effect). Both the processing cost and preferential attention accounts make the same prediction for this contrast, but posit different reasons for the effect. On the processing cost account, scenes showing an avatar with a perspective shared with the participant do not impose a processing cost, while scenes in which the avatar's perspective is not shared do impose a cost since the inconsistent perspectives must be represented and selected between. On the preferential attention account, scenes showing an avatar with an unshared perspective mark some of the balls (those in the joint visual field) as being more salient, and others (those outside the joint visual field) as being less salient, resulting in a delay in account for these balls. Scenes showing an avatar with a shared perspective, on the other hand, contain all the balls within the joint visual field, and therefore do not incur the delay in accounting for balls outside of this field.
2. The preferential attention account predicts that *Double unshared* scenes will be faster than *Single unshared* scenes. This is because *Double unshared* scenes, in which two avatars appear and each sees a unique cluster of balls, do not highlight a more salient group of balls over a less salient group – there is no joint visual field shared between the avatars and

participant, and there is therefore no especially salient cluster of balls driving preferential attention to this cluster and delayed processing of balls outside this cluster. *Single unshared* scenes, on the other hand, do have a joint visual field between the avatar and participant, and therefore have a more salient cluster of balls, driving delayed processing of balls outside this cluster. The processing cost account predicts the reverse: that *Double unshared* scenes will be slower than *Single unshared* scenes, because there are more perspectives to process in these scenes.

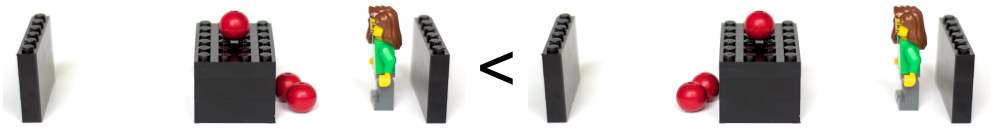
3. The preferential attention account predicts that *Double unshared* scenes will not be significantly different from *Single shared* scenes. In *Double unshared* scenes, where each avatar sees a unique cluster of balls, there is no group of balls in a joint visual field shared between avatars and participant, and therefore no salient cluster of balls driving preferential attention. *Single shared* scenes do have a cluster in joint visual field, but no balls outside this cluster, and therefore no delay in processing the balls located outside of the joint visual field. Neither of these scenes should incur the delay associated with accounting for balls outside of a marked, salient joint visual field, and should therefore have similar reaction times. The processing cost account predicts that *Double unshared* scenes will be slower than *Single shared* scenes, as the *Double unshared* scenes include additional inconsistent perspectives to process and select between (each avatar sees only a subset of the total number of balls in these scenes, making each of their perspectives inconsistent with that of the participant), and should therefore impose a processing cost.

6.2.1 Materials and methods

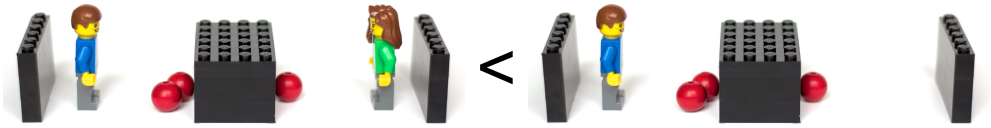
The stimuli from the experiments described in Chapters 4 and 5 were used, with one difference: this expanded set of stimuli included images with both Sally and Andrew present simultaneously, with participants required to take either their own perspective, Sally's perspective, or Andrew's perspective in any given trial (cued by YOU/HE/SHE).

Participants.

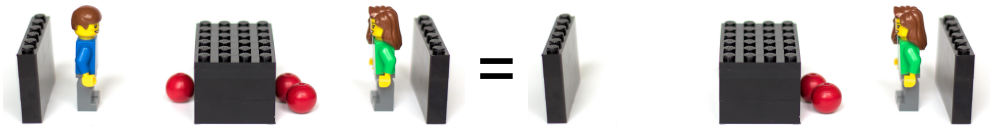
Thirty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 30 minutes.



(a) Prediction 1: *Single shared* scenes will be faster than *Single unshared* scenes.



(b) *Double unshared* scenes will be faster than *Single unshared* scenes.



(c) *Double unshared* scenes will not be significantly different from *Single shared* scenes.

Figure 6.2: The predictions of the dual-avatar task.

Procedure.

Participants completed a short training sequence explaining the instructions for the task, and 16 practice trials with correct/incorrect feedback on responses. This was followed by 384 trials, divided into four blocks, with a self-paced break between blocks. Participants were instructed to judge whether the digit presented before each trial matched the number of balls either they or the selected avatar could see in the picture that followed (cued by YOU/HE/SHE), making this an explicit task.

Because only certain combinations of balls and avatars were of interest to the hypothesis, trials were considered to be either "critical" or "distractor". Critical trials consisted of scenes that showed:

1. One avatar ("single" scene), with a perspective shared with the participant (*Single shared*). This is the same layout as *Avatar sees* in previous tasks.
2. One avatar, not sharing the participant's perspective, and all balls clustered around the central table to control for spatial distribution (*Single unshared*). This is the same layout as *Avatar faces* in previous tasks.
3. Two avatars, with neither avatar sharing the participant's perspective or the perspective of the avatar, with all balls clustered around the central table (*Double unshared*).

The "Consistency" variable in this task is referred to as consisting of "shared" and "unshared" levels in order to make the meanings clearer than the standard "consistent" and "inconsistent" terms, given the range of perspectives in each scene. That is, on an "unshared" scene, the avatars do not share a perspective with each other, and neither shares the participant's perspective; this terminology makes the three disparate perspectives clearer than "inconsistent", which could be taken to refer only the difference in perspective between the participant and either/both of the avatars.

All critical trials were Self trials (those requiring the participant to respond based on his or her own perspective). Because *Double unshared* scenes required either three or four balls to be present, the *Single shared* and *Single unshared* scenes used in critical trials also had either three or four balls. Each of these conditions appeared 32 times (16 times with three balls, and 16 with four) to create 96 critical trials. As there were limited numbers of possible scenes matching these criteria that could be generated from the stimuli, this required repetitions of each image, up to a maximum of eight repetitions.

In order to disguise the repetition of these critical scenes, a further 288 distractor scenes (making critical trials a quarter of the total number of trials) were used in order to create a total trial set that was balanced on every condition likely to be noticed by the participants:

- One vs two avatars;
- Sally vs Andrew (in Single scenes);
- Self vs Other trials;
- In Other, two-avatar trials, the requirement to take either Sally's or Andrew's perspective;
- Shared vs Unshared perspectives;
- Number of balls (between one and four in each scene, with up to two balls in a given slot);
- Yes vs No response.

Sally and Andrew could appear on either the left- or right-hand side of the screen randomly. Because previous analyses showed no significant differences between Yes, No-none and No-other responses (see Chapter 4), half of the trials were Yes and half No, with No trials randomly set to No-none and No-other. In a post-experiment questionnaire, participants were asked whether they noticed any patterns in the questions they were asked, and none reported noticing a high level of repetition of particular scenes.

The experiment was implemented using PsychoPy (Peirce 2010).

6.2.2 Results

We removed training trials and timed-out trials (0.73%, $n = 87$). No trimming was conducted on higher reaction times, given the imposed cut-off of 2,000 ms on all trials. As per Whelan (2008), trials in which the response RT was lower than 100 ms were also removed, on the assumption that these trials could not be genuine responses to the stimuli (0.19%, $n = 22$). Incorrect responses (5.84%, $n = 687$) were removed and not analysed given previous null results of analyses of accuracy. Visual inspection of the reaction time data revealed an obvious deviation from the normal distribution, necessitating a log transform of the data (Baayen and Milin 2010). As with previous analyses, log-transformed RTs were used for the analyses, but we report slope estimates in milliseconds, and plot raw RT means, for the sake of clarity.

The dataset was limited to just the critical trials. We used `lme4` (Bates et al. 2015) and `afex` (Singmann et al. 2017) to perform a series of mixed effects regression analyses on the log-transformed reaction times (logRT). We used the standard $p < .05$ criterion for determining where effects were significant, with p-values obtained using model comparison (likelihood ratio tests) using the `mixed()` function in the `afex` package (Singmann et al. 2017) in R (R Core Team 2015).

We first conducted a linear regression analysis of the altercentric effect; that is, a comparison of RT for *Single shared* and *Single unshared* trials. Consistency was sum-coded, and entered as a fixed effect into the model, with random intercepts for participants and images, as well as by-participant random slopes for the effect of Consistency.[1]

The model showed no effect of Consistency (Table 6.1). This result fails to replicate the altercentric effect found in Chapter 5, Experiment 1, despite this being an explicit task.

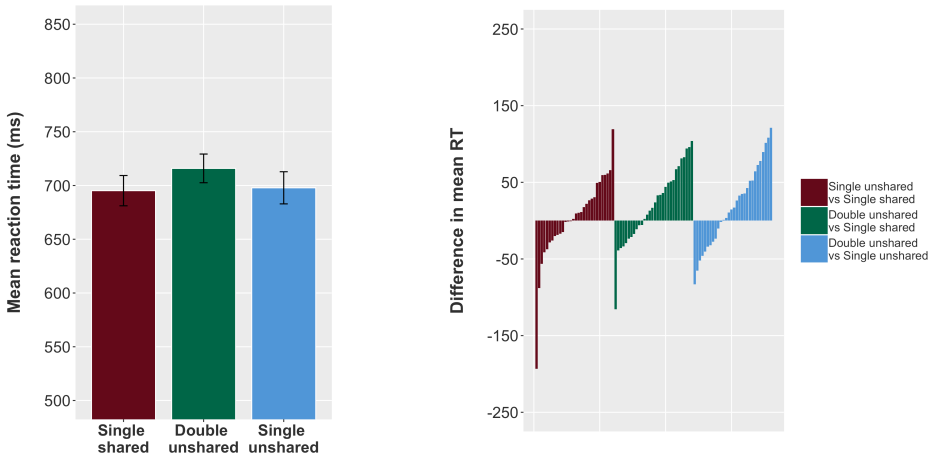
A second analysis modelled the effects on RT of *Double unshared* trials compared to *Single shared* and *Single unshared*. Consistency was sum-coded to create two slopes, one comparing *Double unshared* to *Single shared* and another comparing *Double unshared* to *Single unshared*. Consistency was entered as a fixed effect into the model, with random intercepts for participants and images, as well as by-participant random slopes for the effect of Consistency.[2]

The model showed no effect of consistency, suggesting no significant difference between *Double unshared* trials and either *Single shared* or *Single unshared* trials (Figure 6.3).

[1] Model syntax: `logRT ~ Consistency + (1+Consistency|Participant) + (1|Image)`
[2] Model syntax: `logRT ~ Consistency + (1+Consistency|Participant) + (1|Image)`

Table 6.1: RT across consistency for arrows and avatars.

Slope	β	SE	χ^2	df	p
Single shared vs *Single unshared*	0.0006	0.008	0.008	1	.93
Double unshared vs *Single shared*	–0.010	0.008	4.86	2	.09
Double unshared vs *Single unshared*	–0.008	0.008	4.86	2	.09



(a) (b)

Figure 6.3: Consistency effects for single and double scenes. (a) Mean RT for *Single shared*, *Double unshared*, and *Single unshared* conditions; error bars indicate 95% CIs on the mean of the by-participant means. No comparison was significant. (b) Each individual participant's difference scores between double and single scenes. No comparison was significant.

6.2.3 Discussion

This task did not replicate the altercentric effect found in an explicit task in Chapter 5, and therefore did not support either the perspective-taking or the preferential attention hypothesis for the altercentric effect. The lack of replication of the altercentric effect raises the possibility that in a more challenging task that requires switching between three perspectives, no avatar perspectives are taken spontaneously.

However, although the results of the likelihood ratio tests did not meet the standard $p < .05$ criterion for significance, it did meet the less stringent $p < .10$ criterion, suggesting that replications and adaptations of this task may be useful. Additionally, a slight majority of participants showed a tendency towards slower RTs in Double compared to Single scenes. This task was substantially more challenging than previous tasks, which may have made it less likely that multiple perspectives could be held in working memory simultaneously, explaining the lack of altercentric effect. For this reason, the task was adapted to make it less challenging, and the experiment was repeated.

6.3 Experiment 2

Two changes were made to reduce the difficulty of Experiment 1: the overall number of trials was reduced by 25% to create a shorter task; and Sally and Andrew were anchored to a particular side of the screen rather than appearing randomly on each side, making it easier for participants to identify which perspective would correspond to HE vs SHE instructions on Other trials.

6.3.1 Materials and methods

Participants.

Thirty participants were recruited through the University of Edinburgh Student and Graduate Employment Service. They were compensated £4 for their participation, which lasted approximately 25 minutes.

Procedure.

The instructions and trial sequence were identical to Experiment 1. The same 96 critical trials were used, but they made up a third, rather than a quarter, of all trials, with 192 distractor trials and a total of 288 trials. Balancing procedures were the same as in Experiment 1, and no participants reported noticing a high level of repetition in the post-experiment questionnaire. The experiment was implemented using PsychoPy (Peirce 2010).

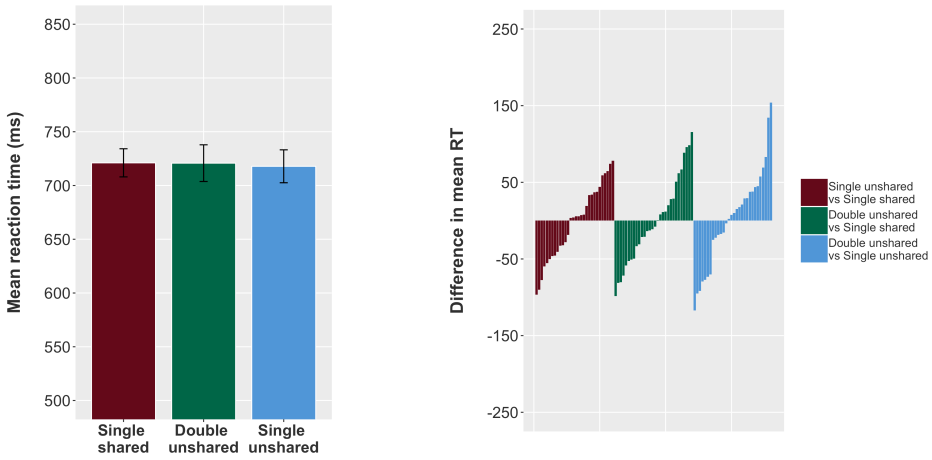
6.3.2 Results

We removed training trials and timed-out trials (1.18%, $n = 99$). No trimming was conducted on higher reaction times, given the imposed cut-off of 2,000 ms on all trials. There were no trials in which the RT was lower than 100 ms. Incorrect responses (2.94%, $n = 244$) were removed and not analysed given previous null results of analyses of accuracy.

The same data transformation, subsetting and analysis procedures were used as for Experiment 1. The models showed no effect of Consistency (Table 6.2), for either the altercentric effect, or for comparisons between *Double unshared* trials and *Single shared*/*Single unshared* trials (Figure 6.4).

Table 6.2: RT across consistency for arrows and avatars.

Slope	β	SE	χ^2	df	p
Single shared vs *Single unshared*	–0.004	0.006	0.462	1	.50
Double unshared vs *Single shared*	0.003	0.007	0.51	2	.78
Double unshared vs *Single unshared*	–0.005	0.008	0.51	2	.78



(a) (b)

Figure 6.4: Consistency effects for single and double scenes. (a) Mean RT for *Single shared*, *Double unshared*, and *Single unshared* conditions; error bars indicate 95% CIs on the mean of the by-participant means. No comparison was significant. (b) Each individual participant's difference scores between double and single scenes. No comparison was significant.

6.3.3 Discussion

This task did not replicate the altercentric effect found in an explicit task in Chapter 5, despite explicitly requiring participants to take the perspectives of the avatars throughout. Given a lack of difference between any *Single* and *Double* trials, this task also provided no evidence for or against either the perspective-taking or the preferential attention hypothesis of the altercentric effect. It is plausible that while the perspective of one other agent can be calculated and sustained throughout a task, the perspective of two different agents overwhelms resources and working memory, and produces no spontaneous perspective-taking. This finding is in line with other tasks finding that involuntary belief tracking is impeded by overall cognitive load (Schneider et al. 2014; Schneider et al. 2012).

However, these results were inconsistent with two previous tasks using dual avatars: unlike Capozzi et al. (2014), there was no altercentric effect found on single-avatar trials; and unlike Michael et al. (2018), there was no evidence of a difference in efficiency in either direction on dual-avatar trials. It is possible that these differences are due to differences in task design and

demand. Unlike Capozzi et al. (2014), this task required taking the perspective of each avatar separately, likely increasing the task demands substantially. In a similar vein, Michael et al. (2018) did not require participants to take the avatars' perspectives at all; and as the previous chapters have demonstrated, direct comparisons of effects in implicit/uncued and explicit tasks may be challenging. Given the surprising effects of stimulus type (Lego vs original avatar) reported in Chapter 5 (Experiments 3, 4 and 5), the differences in avatar/Lego character appearance between this task and previous research introduce a further potentially meaningful difference. Although the explicit task using Lego characters presented in Chapter 5 did find an altercentric effect, there may be limits to how demanding a task with these characters can be before overwhelming participants' tendency to hold the avatars' perspective in working memory throughout.

Given the high number of null results found using the Lego DPT, as well as inconsistent findings across the DPT literature, another possibility is that the altercentric effect may not be as robust as a survey of the literature would suggest. The following section discusses the context in which a seemingly robust result may be more unreliable than it seems, and how this context applies to the DPT literature.

6.4 Chapter summary

This chapter presented two new experiments designed to test the predictions of two different explanations for the altercentric effect: processing cost and preferential attention. Neither task found an altercentric effect, and neither task found differences between dual-avatar and single-avatar scenes, providing support for neither the processing cost nor preferential attention accounts of the altercentric effect. These results contradict with previous studies using dual-avatar designs (Furlanetto et al. 2013; Michael et al. 2018) that have found an altercentric effect on an explicit dual-avatar task, and reduced efficiency on dual-avatar compared to single-avatar trials.

The increased difficulty of the two experiments presented in this chapter may explain this difference. The requirement to switch between three, rather than two, perspectives (including the participant's own perspective) in a scene of substantially increased visual complexity may reduce the likelihood of spontaneous perspective-taking. Simpler tasks may make it feasible to hold the alternative perspective in working memory throughout the task. In a more demanding task such as this one, working memory may be taxed to the extent that it no longer becomes possible to sustain the alternative perspective throughout the task. Because of this, the altercentric interference effect may not be found in more demanding tasks. This finding suggests

that spontaneous Level 1 perspective-taking may be limited to the perspective of only one agent in a sustained task.

However, there is an alternative explanation for these null findings, as well as those presented in Chapters 4 and 5. A broad range of literature has described research practices, incentive structures, and systemic biases in academic publishing that combine to undermine the reliability of the published literature. Because of these factors, effect sizes and false-positive rates are inflated while null results remain unpublished, creating a misleading picture of the reality of a body of research. These problems may be exacerbated in fields that rely on conceptual, rather than direct, replications, since it may be impossible to establish when null results are found because of meaningful differences in task design, and when they are failed replications of false positive effects. The following chapter discusses these concerns.

Chapter 7

The replication crisis and the Dot Perspective Task

The altercentric effect appears to be a robust, widely-replicated result, found successfully even in work that is sceptical of the proposed cause of the effect (Santiesteban et al. 2014; Conway et al. 2017). However, on closer inspection, the effect is not as reliable as it appears to be. Small changes to the experiment design – such as the use of avatars with a different appearance, the pre-trial instructions, and the method of avatar blinding – lead to marked differences in results and, in some cases, a failure to find the key altercentric effect. There are two possible explanations for these null results: first, a lack of understanding of the altercentric effect and the conditions that produce it; and second, that the effect is not as robust or reliable as it appears to be – that is, the altercentric effect, or some instantiations of it, may be a false positive. A combination of these two explanations is also possible: for instance, the altercentric effect may be robust on explicit tasks (as suggested by the finding of an effect in the explicit condition in Chapter 5), but less robust on implicit tasks.

Many published findings in psychology cannot be replicated, with estimates for replication rates ranging between approximately one-third of 100 papers, to two-thirds of 21 papers in a range of more detailed replications (Open Science Collaboration 2015; Camerer et al. 2018). There is a range of possible explanations for these failed replications, including inadequate sample sizes leading to inflated effect sizes and a high rate of false positives; high levels of flexibility in statistical analysis, combined with selective reporting of results; publication bias, resulting in the burying of null results; perverse incentives rewarding questionable research practices and closed data; and small but crucial differences in experiment design. In light of this, is it is worth considering whether the published literature on the DPT may be systematically flawed in line

with some, or all, of these parameters. This is a possibility of some concern, as the DPT has already been used as as indicative of perspective-taking in contexts extending beyond basic research on theory of mind, including research on psychopathy (Drayton et al. 2018); gender, age, and group identity differences (Yue et al. 2017; Schneider et al. 2018; Ferguson et al. 2018); and autism (Schwarzkopf et al. 2014). This broader research context suggests that this is no longer in the realm of pure science, and has the potential for impact beyond basic research.

This chapter reviews some of the major causes of the current replication crisis, identifying how these features pertain to the DPT literature, and makes suggestions for how to proceed in building a more reliable set of empirical results.

7.1 Sample size and statistical power

Null-hypothesis significance testing indicates the probability that a given pattern of results would appear if the null hypothesis were true. If that probability (i.e. *p*-value) is low, this "statistically significant" effect lends support to the hypothesis in question. If the probability of finding the given results under the null hypothesis is high – the standard threshold is a probability greater than 5% – then this is generally taken as providing insufficient evidence to reject the null hypothesis.

A statistically significant finding may represent a true positive: that is, the effect is real, and the study designed to detect the effect did indeed detect it. It may, however, instead reflect a false positive: that is, there is no real effect, but the study detected a pattern of results that has a low probability of appearing under the null hypothesis, suggesting that there are grounds to reject the null. Small sample sizes increase the risk of false positives, due to sampling error – that is, the sampled population is not representative of the overall population, creating the low-probability pattern of results that appears to provide support for the hypothesis (Schmidt 1996).

The inflation of the false positive rate in studies with low statistical power is a result of inflation of the effect size, which is a measure of the strength of a relationship between two variables (e.g. *r*) or the difference between groups (e.g. *d*) (Ferguson 2009). Although there is a "true" effect size that would in theory be detectable in a perfectly constructed experiment with adequate statistical power to detect that effect, the effect size that is actually established in an experiment can be affected by the measurement and by sampling error: if a sample is small, or non-random, it can bias the effect size that is calculated from the data. For instance, a study describing the difference in height between male and female 18-year-olds might conclude that males are, on

average, 60 cm taller than females if it relies on a sample of eight individuals that by random chance happens to include mostly very tall males and very short females. A larger sample size therefore makes it possible to detect the true effect size more accurately (Wetzels et al. 2011).

Sampling error means that the effect sizes that are detected in experiments with small sample sizes can vary widely, making it difficult to estimate the true effect size and also leading to a high rate of false negatives – that is, cases where there is a real effect, but the study does not detect it. Even in cases where a result is a true positive, the effect size in a small sample will, by necessity, be inflated larger than the true effect size: with a small sample size, a larger effect size is needed for the *p*-value to be statistically significant (Schmidt 1996; Ioannidis 2008).

Literature that has a high proportion of experiments with small sample sizes is therefore likely to have a high rate of inflated effect sizes, leading to a high rate of false positives (Cohen 1990). This result is not just the logical consequence of the statistical effects of low sample sizes, but also has support from both empirical and simulation research: across different disciplines, effect sizes that are large in early research are reported to be smaller as research progresses; and simulations of a bionomial distribution, with participants drawn randomly to make up a range of sample sizes, show that studies with inadequate statistical power to detect a given effect size will only pass the threshold of significance by overestimating the effect size (Ioannidis 2008). The inflation of effect sizes makes accurate power calculations for future research difficult, compounding the problem, and is also likely to lead to poor replication rates (Button et al. 2013). The problem resulting from small studies attempting to detect small effect sizes may be so pervasive that simulations of the problem have led to the claim that "most published research findings are false" – due not only to inadequate sample sizes, but also problems such as conflicts of interest and overt chasing of statistical significance (Ioannidis 2005).

Unfortunately, the psychological literature is characterised by "dramatically inadequate" sample size (Marszalek et al. 2011), with no apparent change over the second half of the twentieth century. Median sample sizes in 1951, 1977, 1995 and 2006 ranged between 32 and 60 across different sub-disciplines, with no significant differences between sample sizes surveyed across these years overall (although there were differences in some fields) (Marszalek et al. 2011). Most notably, new guidelines for sample size calculation published in 1999 by the American Psychological Association (Wilkinson and Task Force on Statistical Inference 1999) did not appear to affect sample sizes by 2006. A more recent assessment of 3,801 psychology and cognitive neuroscience papers published between 2011 and 2014 found no improvement in statistical power over the last 50 years, concluding that "false report probability is likely to exceed 50% for the

whole literature" (Szucs and Ioannidis 2017).

Sample sizes in the DPT thus far have been very low, generally below 30 and frequently below 20. The research presented in this thesis is not exempt from this criticism, with only 30 participants per condition. This was a result of resource limitations, as well as simulations calculating that this sample would be adequate to detect the expected effect size with greater than 80% power – an estimate that may have been inaccurate if the assumed effect size was inflated. An important consideration is that the majority of DPT experiments rely on within-subjects designs with repeated measures (that is, the same individual responding to both conditions a number of times – often tens or even hundreds of times), which raises the statistical power substantially (Guo et al. 2013). However, there are designs with between-subjects comparisons, where statistical power will be substantially lower, and statistical power calculations are seldom reported.

Even in the case of within-subjects designs, sampling error may still cause erratic results due to a lack of generalisability. Individual differences in response to the task were often considerable, and effects appear to be driven not – or not only – by a larger effect across all participants, but by a majority of participants experiencing that effect (see Figures 7.1 and 7.2). This may be a result of different individuals interpreting the task purpose and requirements differently, and possibly using a range of strategies to complete the task quickly and accurately, with some individuals or strategies appearing to be more susceptible to the altercentric effect than others. Sampling error, with certain susceptible or non-susceptible individuals making up a disproportionate share of a random small sample, may therefore bias the results, inflating the risk of both false positives and false negatives. This implies not only a risk of false positives and false negatives in conditions predicted to induce the altercentric effect, but also a risk of false positives and false negatives in control conditions such as non-social stimuli or occlusion conditions. The erratic effects necessarily produced by sampling error highlight the need for substantially better-powered research in the DPT.

The persistence of small samples may be a result of lack of resources, a lack of awareness, or the perceived difficulty of power analysis (Maxwell 2004). It may also be a result of the fact that even small sample sizes may yield statistically significant – and therefore publishable – results when multiple comparisons are conducted in the analysis, alongside further options for flexibility, such as data exclusions and a range of dependent variables that may be included or excluded depending on the results (Maxwell 2004). The following section explores the problem of flexibility in analyses.

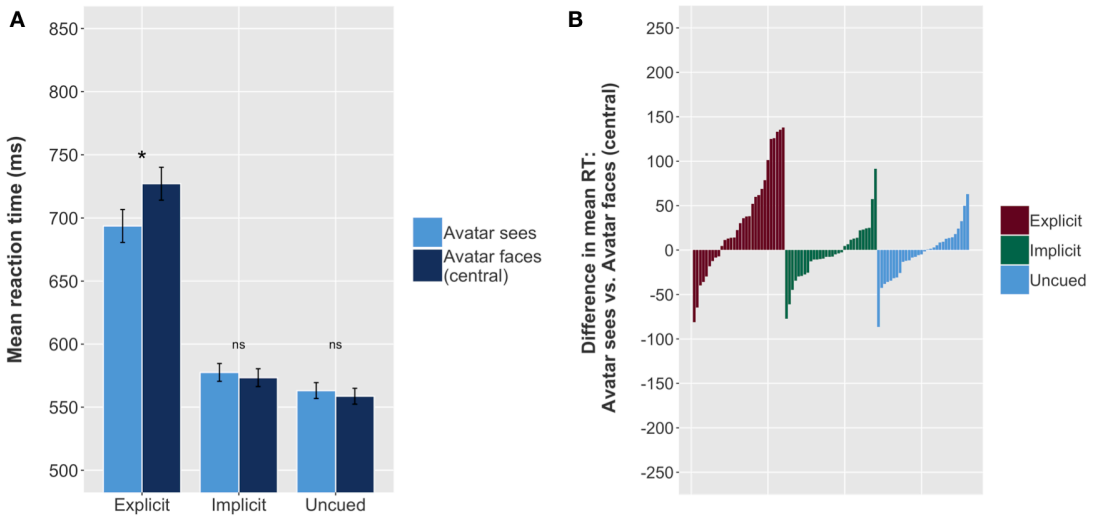


Figure 7.1: Reproduction of Chapter 5, Figure 6 (Effects of Experiment 1, Model 2: *Avatar sees* vs *Avatar faces (central)*). Note that comparing only the means for consistent and inconsistent conditions (A) obscures the considerable differences in individual responses (B). In the implicit and uncued conditions, participant responses range from sizeable differences in the direction of an altercentric effect to sizeable differences in the opposite direction. Even in the explicit condition, a number of participants show a response that runs in the opposite direction to the overall effect.

7.2 Researcher degrees of freedom and the garden of forking paths

Small sample sizes interact with flexibility in analysis to exacerbate the problem of false positives (Maxwell 2004; Ioannidis 2008; Ioannidis 2005). Any given analysis is rife with decisions that need to be made, at every stage: when to stop data collection, which data points should be excluded, what conditions should be compared and with which control variables, whether any data transformations are required, and which dependent variable is the best measure to use. Exploring different alternatives in the analysis, guided by plotting the data, is standard practice, and it is common to report only the analysis that resulted in statistical significance (Simmons et al. 2011).

The nominal rate of false positives across the literature is 5%, based on the standard cutoff for statistical significance: that is, the probability of published results being significant if the null hypothesis is true is less than 5%. However, if multiple analyses are conducted on the same data, the likelihood of at least one of them producing a false positive is higher than 5%.

A series of simulations (Simmons et al. 2011) modelled the effects of freedom in choice of sample size, dependent variables, covariates, and reporting of subsets of conditions, by generat-

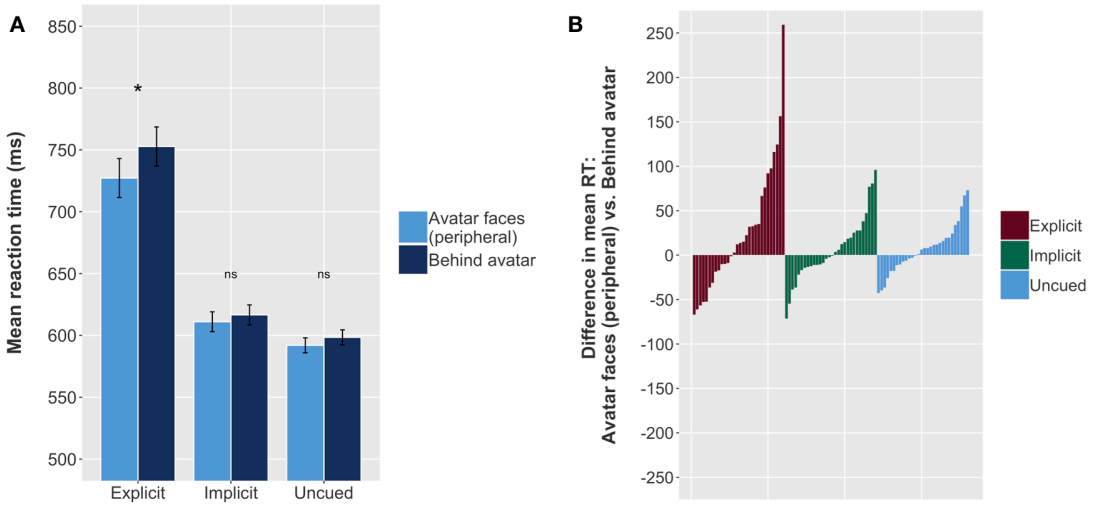


Figure 7.2: Reproduction of Chapter 5, Figure 7 (Effects of Experiment 1, Model 3: *Behind avatar* vs *Avatar faces (peripheral)*). Note that in the explicit condition, the altercentric effect was significant in pairwise comparisons (A), despite a lack of significant effect in the omnibus model. The differences in individual responses (B) suggest that this result was driven not, or not only, by a larger average effect across all participants in the explicit condition, but by a majority of participants demonstrating an altercentric effect, and in a limited number of cases, a sizeable one.

ing random samples from a normal distribution, analysing each sample with various parameters changed in the analysis, and recording the number of results that met the threshold for statistical significance. Flexibility in analysis substantially inflated the probability of finding a significant effect; for instance, analysing two dependent variables instead of one raised the likelihood of finding a significant effect to 9.5%. This is worth considering in light of the use of both RT and error rates in many DPT analyses (as well as IES in some tasks), with some tasks finding effects in one but not the other, most notably a failed replication of Furlanetto et al. (2016) that found an effect on error rates, but not RT (Marshall et al. 2018).

In these simulations, a combination of multiple common techniques, including optional stopping (gathering data only until the point of finding statistical significance is achieved), multiple comparisons, and controlling for variables such as gender when there is no specific hypothesis regarding the role of the control variable, led to as high as a 61% false positive rate – that is, a higher probability of finding a false positive effect than not. Since there are additional possibilities for flexibility in analysis, including more than two dependent variables, and variable participant or data exclusion practices, this may be a conservative estimate for certain cases (Simmons et al. 2011).

These "researcher degrees of freedom" allow studies conducted with small samples to nonethe-

less yield statistically significant results – but the pattern of *which* results are significant may vary between studies, creating contradictions across the literature (Maxwell 2004). Flexible analyses also result in inflated effect sizes, resulting in another layer of inaccuracy across the literature if only the highest effect sizes are reported (Ioannidis 2008).

On its own, flexibility would not increase effect size estimates or false positive rates if all analytical options were presented with no preference; however, this is not the norm. One set of choices is standardly published, and it is unclear whether this is the result of a random choice, or the presentation of the analytical choices with the most favourable outcome. A survey of published literature from a grant program that required materials to be published found that 70% of studies did not report all the outcome variables assessed using the study materials, and 40% of studies did not report experimental conditions (Franco et al. 2016). The effect sizes that were reported were approximately double the effect sizes that remained unreported, and were approximately three times more likely to pass the significance threshold.

This problem is described as arising from repeated analyses used to achieve the desired results (i.e. statistical significance) (Wagenmakers et al. 2012), but a crucial insight is that the "garden of forking paths" can lead an analysis astray even when there is no so-called "p-hacking" (Gelman and Loken 2013). When there is a high number of potential comparisons, a single analysis may use only one of these possible sets of choices, but still result in inaccurate effect sizes if that analysis is guided by the data that resulted from the task. That is, if researchers perform the single analysis that seems reasonable based on theoretical considerations and the dataset available, but would have performed a different analysis had the data been different (that is, with different decisions on questions like data exclusions and transformations), the problem of multiple comparisons persists. With noisy data, small effect sizes, and low statistical power, the analysis that seems reasonable might vary substantially from one dataset to the next. This means that multiple *possible* comparisons in a data-led analysis is enough to introduce the problem, without ever embarking on multiple *real* comparisons.

This is illustrated by a study that required 29 teams, with 61 different data analysts, to analyse the same dataset to address the question of whether dark-skinned soccer players receive more red cards than light-skinned players (Silberzahn et al. 2018). The effect sizes found by the 29 different analyses varied widely, 69% of the teams found a statistically significant effect, and there were 21 unique combinations of covariates used across the analyses.

Even without intentionally hunting for statistically significant effects, potential multiple comparisons may lead to a single analysis producing a result that is not reflective of reality. The

range of conclusions that could be reached depending on the analysis used was also demonstrated in recent work implementing "specification curve analysis", which reports the results of all possible analysis choices for a given question (in this case, the potential negative effects of digital technology on adolescents' emotional wellbeing). In some cases, there were thousands of possible analyses for a single simple research question, and the range of findings was broad enough to encompass polar opposite conclusions that were nonetheless supported by data (Orben and Przybylski 2019).

This work demonstrates not only the danger of multiple possible comparisons, but also how misleading the interpretation of results can be if the analysis is guided by the data. If a hypothesis is formed on the basis of the results, and presented as an *a priori* hypothesis that informed the experiment design – a practice dubbed HARKing, or hypothesising after the results are known (Kerr 1998) – an essentially spurious result may easily appear convincing and robust (Maxwell 2004).

The DPT literature displays many of the characteristics identified as risk factors for a high rate of false positives: possibly insufficient statistical power (in the case of the DPT, most likely leading to sampling error); a high number of tested relationships with no prespecification of what tests will be conducted; a high degree of flexibility in experiment design and analysis, in a field that prioritises statistical significance, (Ioannidis 2005); a lack of consensus on which analytic practices should be default (Ioannidis 2008); and the potential for researchers who may "fervently support their line of research and beliefs" (Ioannidis 2008).

The DPT is, on the face of it, a very simple task, but the analysis offers boundless opportunities for researcher degrees of freedom. A standard analysis may involve two or even three dependent variables (RT, errors, IES). Each of these DVs is likely to be analysed using at least a two-by-two ANOVA comparing the DV on inconsistent and inconsistent trials, as well as some other variable (stimulus type, or sightedness of the avatar, etc.). The RT data may be analysed with the erroneous trials included, or excluded; No trials may be included or excluded; and there are no conventions on data cleaning processes such as the removal of outliers. In one case, both frequentist and Bayesian analysis techniques are used (Conway et al. 2017), a practice that has been identified as presenting a further degree of researcher freedom (Simmons et al. 2011). This "garden of forking paths" provides ample opportunity for analyses guided by the data, and selective reporting of analyses.

7.3 Publication bias

The likely proliferation of false positives across the literature is compounded by the "file drawer" problem: positive results are vastly more likely to be published than null results, which may create a heavily misleading picture over time. Publication bias, combined with a high false positive rate, may create a situation in which no or few null results are known to exist, while many or all of the positive results that have been published are spurious (Rosenthal 1979). This bias may function through researchers neglecting to submit null results on the assumption that they will not be accepted for publication, as well as through reviewers and journal editors declining to publish those null results that are submitted (Sterling et al. 1995).

The percentage of unpublished studies in psychology has been estimated to be around 50% of the research conducted, based on assessments of planned studies that were not published (Cooper et al. 1997; Coursol and Wagner 1986; Shadish et al. 1989). Bias is also reflected in the finding that null results are rare across the literature: less than 10% of published papers assessed in 1995 reported a null finding, despite 30 years having elapsed since publication bias was first recognised (Sterling et al. 1995). A more recent analysis estimated that over 80% of publications across all fields report positive results, with the highest proportion of positive results in psychology (Fanelli 2010). An estimate of the proportion of significant, positive results across the psychology literature suggests that this proportion is approaching 100% (Ioannidis 2012). Meta-analyses have found further evidence of publication bias in psychology (Ferguson and Brannick 2012; Francis 2012b).

Another high estimate places the rate of positive results across the literature at around 96%, pointing to the use of multiple small, underpowered studies as a route to almost-certain statistical significance. Compared to a single, powerful study, this strategy is much more likely to result in "publishable" research, as demonstrated by simulations of a variety of research strategies (Bakker et al. 2012). Where a single high-power study is a high-risk use of resources, small studies that can be analysed with a high level of flexibility are more likely to produce results.

Given an incentive structure in which a high number of publications is associated with career success, and publication is dependent on positive results (Vankov et al. 2014), this is the most rational strategy to pursue. An optimality model exploring the most rational research strategy given current incentives predicts that the most rational strategy is to conduct mainly small studies with low statistical power, seeking novel results rather than powerful studies confirming or replicating existing tentative results (Higginson and Munafò 2016). Publication bias therefore

creates an incentive structure that encourages the use of the analytic techniques discussed in the previous two sections that inflate the false positive rate.

In order for such high rates of positive results to be reflective of reality, all the studies conducted would need to have had high statistical power, and researchers would need to investigate only true hypotheses. Clearly, neither of these can be the case, and effect sizes found in theses and other unpublished work are smaller than those in published literature, suggesting that published papers are not reflective of the true findings of total research effort (Sterling et al. 1995). Since studies do not have sufficient power to detect the effect in question 100% of the time, there should also be a predictable number of false negative results, i.e. the Type II error rate. Even a long series of successful replications, with no unsuccessful replications, may be an indication of publication bias, if there are no published unsuccessful replications at a rate that would be expected based on the predictable Type II error rate (Francis 2012a).

The DPT has been successfully conceptually replicated multiple times (barring the null result in tasks with 0 ms SOA), and this may indeed be an indication that the altercentric effect is a true effect. However, given the high number of tasks, a certain number of false negatives would be expected at a standard Type II error rate.

A Google Scholar search of papers citing the original DPT (Samson et al. 2010) and eliminating those that do not use some kind of avatar-and-dot-based assessment of mindreading suggests that there are currently 39 published papers using the DPT (many of which include multiple experiments), excluding the work in this thesis. Conservatively, with a Type II error rate of just 5% and assuming only one experiment per paper, this suggests that there should be at least two published failed replications (again, not counting the results in this thesis). While there are published versions of the task not finding the altercentric effect, these cannot strictly be considered failed replications, since these findings have largely been explained by methodological differences. Given the small sample sizes used, this conservative rate of 5% is likely to be far too low. The lack of published null results is therefore some indication that publication bias is likely to be at play, suggesting that there may be unpublished failed replications of the DPT.

7.4 Questionable research practices

> *"We knew many researchers – including ourselves – who readily admitted to dropping dependent variables, conditions, or participants to achieve significance. Everyone*

knew it was wrong, but they thought it was wrong the way it is wrong to jaywalk. We decided to write "False-Positive Psychology" when simulations revealed that it was wrong the way it is wrong to rob a bank." – Simmons et al. (2018, p. 255)

Collectively, the "questionable research practices" (QRPs) described above have created a problem with unreliability and lack of reproducibility across the literature not just in psychology, but a wide range of fields, including economics (Camerer et al. 2016), and clinical research (Osherovich 2011; Prinz et al. 2011; Begley and Ellis 2012; Freedman et al. 2015). A 2016 survey of 1,576 researchers found that 90% agreed that there was a reproducibility crisis, with 52% considering it "significant" and 38% considering it "slight". Only 3% said there was no crisis, with 7% answering that they did not know. High numbers of respondents (between 60 and 90%) across a range of disciplines, including physics and chemistry, reported not being able to reproduce another researcher's experiment. Selective reporting, pressure to publish, low statistical power, and poor analyses were highlighted as the most common causes of irreproducibility (Baker 2016).

QRPs are not considered to be research misconduct. An anonymous survey of over 2,000 psychologists about QRPs found that many practices were considered to be defensible in certain contexts, or even common enough to be considered the norm (John, Leslie K et al. 2012). The survey offered an incentive for truth-telling in the form of an increased charitable donation on behalf of each respondent if an algorithm predicted a high degree of truthfulness in that individual's responses. Rates of self-admission for a range of QRPs were high, and in some cases approached 100%. The survey covered practices such as excluding data after exploring the results of doing so; failing to report all of the dependent variables, conditions, or results; viewing results partway through data collection and deciding whether or not to continue based on the outcome; erroneously rounding down p-values; and hypothesising after the results of the study were known. In some contexts, some of these practices may be genuinely defensible (e.g. not reporting results of a dependent variable where the results were redundant with the main DV explored, for the sake of brevity); in others, that is not the case. The moral grey area may be what allows these practices to continue.

Errors and bias introduce further unreliability into the literature beyond QRPs. An analysis of 281 articles in the psychological literature showed that approximately 18% of statistical results were misreported, and that 15% of the articles assessed had at least one error that changed the conclusion of the study (from significant to non-significant, or vice versa), often in line with the expectations of the researchers, suggesting bias-led errors (Bakker and Wicherts 2011). Further evidence corroborates the possibility of such bias-led errors: researchers have been shown to be

subject to confirmation bias in analysing their data, for example by dismissing data that contradicts their expectations as problematic on methodological grounds, while data that is consistent is accepted and not scrutinised to the same extent (Fugelsang et al. 2004).

The lack of availability of data makes it difficult to assess the extent of QRPs in a given paper. One attempt to assess this availability failed to obtain more than 26% of the data sets from 141 articles published by the American Psychological Association, requested by email, despite all being published in a journal that required signing an ethical statement in support of data sharing (Wicherts et al. 2006). An unwillingness to share data was found to be related to weaker evidence and a higher rate of errors in statistical analysis, suggesting that reluctance to share may be related to concerns that additional analyses may find flaws in the analysis and produce contradictory evidence (Wicherts et al. 2011). However, an alternative reason may be the amount of work that is involved in sharing data (Wicherts et al. 2006). Journal policy may go some way towards rectifying this, but this is not necessarily the case: the majority of 351 papers published in journals with a policy on public data availability were found to not adhere fully to the policies, and only 9% of the papers assessed published full, raw data online (Alsheikh-Ali et al. 2011).

7.5 The mechanics of replication

A lack of access to not only data, but also methods, materials, and analysis code make replication attempts difficult (Ioannidis 2012). Over the past century, replication has been rare: since 1900, only around 1% of papers across 100 psychology journals were directly replicated, with replications less likely to be successful when authorship did not overlap between the original and the replication. However, the rate has been increasing over the decades, reaching approximately 2% of all articles in 2012 (Makel et al. 2012). Presumably the rate has continued to increase due to increasing awareness of the replication crisis, although more recent figures are not available. There have, however, been many recent high-profile replication attempts, the results of which suggest a low rate of replicability but also highlight the many difficulties involved in producing a direct replication.

A large replication effort attempted directed replications of 100 psychology studies, evaluating replicability using statistical significance, effect sizes, and subjective assessments of whether a task had replicated. While 97% of the original studies had statistically significant results, only 36% of the replications were significant, suggesting that only approximately one-third replicated with significance as the deciding factor. The replication rate was different for other measures,

however: 47% of the effect sizes from the original studies fell within the 95% confidence interval range of the effect sizes found in the replication; and the subjective assessments suggested that 39% of the originals had replicated successfully (Open Science Collaboration 2015).

This attempt was criticised on various grounds, particularly important differences in methods between the original studies and replications. For instance, one of the replication studies used emails instead of letters in a study on the wording of a charitable appeal. Only 69% of the authors of the original studies had endorsed the replication protocols, suggesting that the tasks used may not have been unambiguously good tests of the hypotheses in question (Gilbert et al. 2016). The role of various changes in methods are clearly highly debatable; a replication attempt that found successful replications for 10 out of 13 studies reported that there were no changes in replication success caused by differences such as using lab vs online samples, or US populations compared to populations in other countries (Klein et al. 2014). Exact replications are also impossible, given that any replication will involve different experimenters, equipment, locations, weather, recent news events, etc.. The changes to protocol that are likely to affect the results in predictable and meaningful ways is likely to be a subject for ongoing debate (Anderson et al. 2016), with no current consensus on what constitutes a successful replication (Brandt et al. 2014).

The inflated effect sizes likely to be present in the original studies may have played a role in the low replication rate (Anderson et al. 2016; Etz and Vandekerckhove 2016; Ioannidis 2012), since the replication attempts used the published effect sizes to determine statistical power. If the effect sizes in the original studies were inflated, the statistical power for the replications would likely have been insufficient to detect the true effect sizes. A later replication attempt of 21 psychology studies in high-prestige journals used sample sizes that were, on average, five times higher than the original studies, and found a higher rate of successful replication: 13 out of the 21 studies replicated, but effect sizes were, on average, 50% of the effect sizes in the original papers (Camerer et al. 2018). These failed replications do not support the conclusion that the effects in question do not exist. Rather, they suggest that there is inadequate evidence for those effects (Simonsohn 2015).

The DPT appears to produce a robust effect given its appearance in a wide range of tasks. However, it is essential to note that the literature consists entirely of conceptual replications, with considerable differences in factors like the appearance of stimuli, task instructions and setup, and experiment design. A high rate of conceptual rather than direct replication may interact with publication bias to create an appearance of a widely-replicated phenomenon that is

in fact less robust than it appears to be. Published direct replications that find null results are rare (Makel et al. 2012; Sterling et al. 1995), but may still be considered worthy of some attention and decrease confidence in the original finding; null conceptual replications, on the other hand, are more likely to be taken as an indication that some subtlety of experiment design has not been implemented properly and is therefore responsible for the null result (Pashler and Harris 2012; Simmons et al. 2011); this may also lead to the conclusion that such subtleties play a meaningful role in an effect, which may be a mistaken conclusion based on insufficient evidence.

Although failed conceptual replications may affect conclusions about how far the original effect can be generalised, they do not usually result in uncertainty about the robustness, size or replicability of that effect. Because failed conceptual replications are taken to be an indication of failure of experiment design, rather than a genuinely failed replication, unsuccessful conceptual replications are unlikely to be published, while successful conceptual replications are likely to be published and provide further evidence of the effect in question. In this way, conceptual replications are subject to, and may even amplify, the effects of publication bias. Contributing to this bias is the possibility that conceptual replications may be more likely to be attempted because their novelty compared to direct replications makes them more likely to be publishable (Pashler and Harris 2012).

Each novel variation of an experiment produces a range of new ambiguities, making it difficult to determine what precisely is responsible for the varying results – a spate of false positives and unpublished true negatives, or some subtlety of the effect that is dependent on variations in the experiment design. This is precisely the difficulty apparent across the DPT literature, with its wide range of experimental and analytical choices. There are differences in what kind of avatar is used; the dimensions of the room; how dots are positioned in the room; the relative sizes of the objects; and the colour, size, and spacing of control objects, all of which should be controlled for in a task that is testing visual processing. There are differences in the instructions given to participants, the length of display of various elements of the trial procedure, and how responses are gathered (via mouse, keyboard, button box, and the sides of Yes and No responses). The experiments have a wide array of designs, with the number of trials, filler or distractor trials, internal balancing of different trial types, and division of trial types into separate blocks all varying widely. Then, as noted in Section 7.2, there are substantial differences and opportunities for flexibility in data analysis practices.

Collectively, this variation makes it impossible to directly compare different DPT experiments, or even to determine what would constitute a failed replication; in combination with

likely publication bias, exacerbated by the proliferation of conceptual replications and a lack of adequately powered direct replications, it is difficult to determine whether the crucial alter-centric effect is replicable, and in what circumstances. As difficult as it may be to determine what constitutes a direct replication, a direct replication that is as close as possible to the original work is necessary to explore the strength of the effect, with further high-powered research implementing controlled changes to the methods necessary to explore the factors that modulate appearance of the effect.

7.6 Proposed solutions

A range of best practices have been proposed to alleviate the problems discussed above and increase the replicability of research. Some authors have advocated the widespread adoption of new statistical practices, such as a norm of thorough reporting of all conditions, variables, and analyses (Simmons et al. 2011); a move away from null-hypothesis significance testing and towards the use of effect sizes and confidence intervals (Sterling et al. 1995; Simonsohn 2015); increased use of Bayesian methods (Etz and Vandekerckhove 2016); and a controversial proposal to change the commonly accepted threshold of significance from $p = .05$ to .005 (Benjamin et al. 2018), which has prompted alternative suggestions to abandon significance testing altogether (McShane et al. 2019; Amrhein and Greenland 2018) or justify the significance threshold that is appropriate to the question (Lakens et al. 2018), on the grounds that a lower threshold would simply encourage more HARKing and selective reporting.

Non-statistical best practice, including blinding experimenters to experimental conditions both in the lab and during analysis; declaration of all conflicts of interest (including non-financial conflicts; Munafò et al. 2017); and open sharing of materials, data, and analysis scripts in trusted repositories are all strongly recommended (Nosek and Lakens 2014; Munafò et al. 2017; Nosek et al. 2015; Maxwell 2004). Further suggestions include meta-analyses, widescale multi-lab collaborations to increase statistical power and consensus on experiment design, and well-powered replication attempts (Maxwell 2004; Nosek et al. 2015).

Changes to research workflow and publication norms have been the subject of widespread discussion, and to some extent, adoption (Kidwell et al. 2016; Nosek et al. 2018; Nosek and Lindsay 2018). Pre-registration, which requires researchers to describe their hypotheses, methods and analyses before gathering data, may be a useful way to prevent partial reporting of conditions and analyses, HARKing, and p-hacking or following a garden of forking paths (van 't Veer

and Giner-Sorolla 2016; Nosek et al. 2015; Wagenmakers et al. 2012; Maxwell 2004).

Pre-registration would also draw a clearer distinction between *confirmatory* research, which sets out to establish whether data supports a pre-established hypothesis, from *exploratory* research, which seeks to observe patterns in the data in order to generate hypotheses, and is a useful contributor to empirical research, but should be labelled as exploratory to distinguish it from the stronger degree of evidence provided by confirmatory research (Wagenmakers et al. 2012; van 't Veer and Giner-Sorolla 2016; Nosek and Lakens 2014). Registering the intention to run an experiment before doing so has the additional benefit of providing some buffer against publication bias, since a public record of the experiment will exist regardless of eventual publication (Wagenmakers et al. 2012; Nosek et al. 2015; Munafò et al. 2017) and should ideally make it easier to identify.

Pre-registrations, however, may be variable in quality, ranging from a brief description of the analysis plan to the full experiment materials and analysis script. There is limited control on the adherence of the paper to the pre-registered plan, with any such control essentially dependent on the initiative of the reviewers. An alternative model is the registered report, which requires peer review before the data is gathered (Sterling et al. 1995; van 't Veer and Giner-Sorolla 2016; Nosek and Lakens 2014; Chambers 2013). This not only provides a more robust control on the research adhering to the pre-specified plan, but also allows for peer review to contribute to the research design before it is implemented (van 't Veer and Giner-Sorolla 2016), and guards against publication bias by committing to publication on the strength of the methods, regardless of the results (Chambers 2013).

7.7 Conclusion: the way forward for the DPT

The replication crisis literature is, to a large extent, focused on social psychology and clinical research, given the high-profile failures of replication in these fields. The lack of research on the specific replication rates of different fields, including cognitive psychology and developmental psychology, should not be taken as an indication that these fields are unlikely to have similar problems; given the common risk factors across all fields that are based on statistical inference, these problems are likely to be shared. There may be different rates of, for example, publication bias, QRPs, and replication success across different disciplines, but the basic principles underlying the problem are the same.

The DPT literature displays the hallmarks of a field of research that is likely to have a high

rate of false positives: small sample sizes (and concomitant likely inflated false positive rate); ample opportunity for flexible statistical analysis; a lack of negative results, suggesting possible publication bias; and limited sharing of materials, analysis scripts and data. A dearth of direct replication attempts, alongside multiple conceptual replications with widely varying experiment design, is of particular cause for concern. The current pattern of inconsistent results across the DPT literature may have both statistical and theoretical explanations; these can only be distinguished by high-quality empirical work, ideally direct replications with high statistical power, published as registered reports.

Although considerable efforts have been made to adhere to best practices in the research presented in this thesis – with preregistration as well as open data, materials and analysis – this work is not immune to the problems outlined above. The identification of the boundless opportunities in the analysis of the data presented in Chapter 4 is what determined the need for preregistration of the experiments in Chapter 5 and 6. The work in Chapter 4 demonstrates the utility of an exploratory analysis to identify patterns in the data that form the basis of a preregistered confirmatory test. Nonetheless, there is potential improvement possible here, such as a "multiverse analysis" (Steegen et al. 2016) that involves conducting all possible analyses of the data and reports the results of all, thereby determining how robust a significant effect is.

The decision was made to pre-register this work rather than attempt to publish as a registered report because of the limited timeline available for the research; a lengthy peer review process for a registered report may have made it impossible to begin data collection within the necessary timeframe. The pre-registered plan was nonetheless as detailed as possible, including an analysis script that had been written based on simulated data and was therefore used without alterations in the experiment analyses. However, there were minor deviations from the pre-registration, as noted in Chapter 5: the pre-registration did not adequately describe when participants' data would be eliminated (for example, in the case of computer failure), and in error omitted the detail that Other trials would be eliminated from the analysis of data in the explicit condition. Adherence to this preregistered plan, though, was not a condition of publication, making a preregistration a less stringent open science practice than a registered report.

Perhaps most importantly, the sample sizes in this research were low. Given the appearance of substantial individual differences in participant responses, this may have resulted in sampling error. Resources were limited, making it necessary to conduct the research with the smallest feasible sample size, and power calculations indicated that 30 participants per condition would be sufficient for the expected effect size, with greater than 80% power. However, these calculations

were based on assumptions of what the expected effect sizes would be, informed by Samson et al. (2010). A more robust calculation of expected effect size would take into account effect sizes from across the literature, anticipating the possibility that published effect sizes may be inflated (Camerer et al. 2018; Ioannidis 2008) and justify an appropriate significance threshold (Lakens et al. 2018).

Future research on the DPT should aim to adopt the best practices identified in Section 7.6, using larger sample sizes, direct replications, and pre-registered methods and analyses. While large-scale replications of some of the basic findings in the literature (for instance, altercentric effects on explicit tasks) should be given primacy, other fruitful avenues for research and replication include confirmation of the pattern established in Chapter 5 (implicit and explicit), as well as a robust investigation of the differences in Lego vs avatar stimuli that appear to affect DPT results. Opportunities for exploratory research include an investigation of how increasing demands of the task itself, rather than through a simultaneous distractor task, contributes to the altercentric effect; an investigation of the individual differences underlying varying responses to the task; and the effects of various kinds of stimuli, ranging from live humans to non-humanoid cartoon characters, to understand how perceptions of agency affect perspective-taking.

There is an urgent need for research of higher quality on the DPT, given its relevance both to basic research on mindreading, as well as research with a range of clinical and broader applications, including psychopathy (Drayton et al. 2018); group differences (Yue et al. 2017; Schneider et al. 2018; Ferguson et al. 2018); and autism (Schwarzkopf et al. 2014). Clarity is needed on when an altercentric effect is present and what cognitive processes drive that altercentric effect. The DPT clearly has potential to test hypotheses generated by robust theory, and as such produce important findings that will contribute to building a clearer understanding of mindreading. The potential of the paradigm, and its extension beyond basic and into applied research, makes it essential to conduct more robust research and gain a clearer understanding of the altercentric effect and its limitations.

7.8 Chapter summary

This chapter has reviewed the literature on the systemic statistical and institutional factors underlying low replication rates in fields based on statistical inference: low sample size and a concomitant high rate of false positives; researcher degrees of freedom; publication bias; incentivisation of questionable research practices; the difficulty of replication; and a high rate of conceptual

replications. It has discussed the current recommendations of best practices, including open materials, data and analysis; pre-registration and registered reports; increased sample sizes and collaborative research; and a distinction between exploratory and confirmatory research.

It has assessed the DPT literature in terms of these concerns, including the work presented in this chapter, and argued that many of the flaws identified in other fields apply to the DPT. This suggests that the current inconsistent results across the DPT literature may be explained by a high false discovery rate, a poor understanding of the abilities assessed by the task, or both. Finally, suggestions for future research were identified.

Chapter 8

Common ground: reframing ostensive-inferential communication

As discussed in Chapter 3, the DPT is a useful paradigm for testing the predictions of the two-systems and submentalising accounts of mindreading. This is relevant to the ostensive-inferential account of communication, which demands context-based metarepresentations in every communicative interaction, and therefore rests on a one-system, mentalising account of mindreading. Evidence that the rapid and involuntary mindreading found in the DPT is consistent with this one-system, mentalising account is therefore evidence in support of the kind of mindreading required by the ostensive-inferential account.

The evidence presented from the DPT experiments in Chapters 4, 5 and 6 is not consistent with either submentalising or two-systems accounts. As discussed in Chapter 5, the results are best explained by mentalising, not submentalising, suggesting that this form of rapid and involuntary perspective-taking entails representation of the avatar's visual perspective rather than a process of directional orientation. Further, the lack of any effect on a range of tasks – those tasks using Lego avatars and not requiring participants to switch between perspectives (i.e. implicit and uncued tasks); implicit tasks with human avatars that cross a certain threshold of visual complexity; and complex explicit tasks – suggests that the effect is not automatic, but rather spontaneous: that is, it is deployed selectively, when contextual cues are sufficient to draw participants' attention repeatedly to the relevance of the avatar's perspective, and when the demands of the task do not overwhelm cognitive resources to the extent that an alternative perspective can no longer be sustained throughout the task.

Although many of the tasks have not found evidence of an altercentric effect, it should be noted that the explicit tasks *have* found a near-ceiling effect on a task that requires perspective-

taking of another agent, including in tasks that required switching between multiple perspectives (i.e. the experiments presented in Chapter 6). This is a demonstration not necessarily of involuntary or spontaneous perspective-taking, but certainly of rapid perspective-taking.

Given the possible problems with replication in the DPT discussed in Chapter 7, these conclusions should be treated with caution until they are replicated by experiments with greater statistical power, and until the roles of different stimuli and methodological parameters are properly understood. In the absence of stronger contradictory evidence, though, these results offer tentative support for the plausibility of the ostensive-inferential account. The support is tentative not only because of potential empirical weaknesses, but also because communicative intentions require recursive mindreading that includes representation of a wide range of mental states, rather than the more limited non-recursive Level 1 visual perspective-taking assessed in the DPT. Because of this, research on visual perspective-taking may be useful for testing general predictions about the kind of mindreading required by ostension, but cannot test the ostensive-inferential account more directly.

This chapter takes a closer look at the cognitive abilities involved in communicative intentions, and suggests a new avenue for future research on ostension. I argue that communicative intentions are more closely related to joint attention than to mindreading, and that the literature on joint attention offers conceptual tools that solve certain problems with the ostensive-inferential account of communication.

First, I recap the basic principles of communicative intentions, and discuss why this raises concerns about cognitive plausibility. I then introduce Moore's (2014, 2016) minimalist account of communicative intentions (which attempts to address these concerns by presenting a less demanding account of the mindreading involved in communicative intentions) and offer a critique of this account.

In Section 8.2, I characterise joint attention and a range of related concepts, explaining the terminology I will use in this chapter. I use these concepts to present an alternative characterisation of communicative intentions that rests on these concepts rather than mindreading; and I note similarities between this recharacterisation and a range of other related concepts, including common ground and mutual manifestness.

Section 8.3 explores the improvements offered by this recharacterisation: it addresses the critique of cognitive implausibility, and provides a more complete and accurate characterisation of what communicative intentions entail. I describe the problem of infinite regress in joint attention, and following Wilby (2010), I argue that a joint state cannot be reduced to the individual

mental states of the participants, but rather that "second-person" states like this are irreducible and, crucially, do not entail representing another individual's perspective as separate from one's own.

Finally, in 8.4, I suggest that the recharacterisation of communicative intentions opens up an alternative path for future research, with a more constrained and specific role for mindreading. I suggest a range of research questions raised by this alternative characterisation.

8.1 Reducing the metarepresentational demands of Griceanism

In Chapter 2, I introduced the ostensive-inferential account of communication, which characterises human communication as communicative intentions that provide evidence for the intentions of the speaker (Scott-Phillips 2015; Sperber and Origgi 2012). These communicative intentions require recursive mindreading from both the speaker and hearer. To recap the description as laid out in Chapter 2:

> Imagine that my partner Johnny and I realise that we must post a letter the next day, and I agree to do it in the morning. When the morning comes, however, I find I am running late and will not have time to post the letter. Rushing out of the house with no time to write a note, knowing that both of us will be in meetings all day with no time to text or call, and aware that Johnny will be home first, I do the quickest thing I can think of and place the letter prominently on top of his computer keyboard on his desk, where I know he will see it soon after entering the house.
>
> I do this knowing that he will understand the message: "I didn't get around to this; could you please take care of it?" That is, Johnny will understand that the placement of the letter was a communicative behaviour, and will therefore be able to understand that I had an informative intention and work out what the message was by drawing on our shared context – that we'd discussed the previous night that the letter really needed to be posted today; that I'd agreed to do it in the morning, but clearly hadn't; that we'd not had time to talk all day; and that I'd have been aware of him coming home first, and where he would be likely to go first in the flat.
>
> This probably sounds like an unruly amount of information to keep track of. That is a feature, not a bug, of this account, which argues precisely this point: that a communicative gesture like this entails a great deal of both context and mindreading from both the sender and receiver, and yet is somehow accomplished by people

with something close to effortlessness. I would probably not think about it for longer than a couple of seconds if I had to come up with it while rushing to leave the house, and when I tested this hypothetical scenario on real-life Johnny, he decoded it instantly. This, despite the fact that a formal account of a *communicative intention* – that is, an action in which the signaller is attempting to make it clear that their behaviour itself is communicative – requires Johnny to entertain a fourth-order metarepresentation.

To recap, in order to interpret an informative intention, he will need to entertain the second-order metarepresentation that is my informative intention: Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$

My communicative intention is a third-order metarepresentation: I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$

This means that, in order to register my communicative intention – the signallhood of my signal – Johnny must entertain a fourth-order metarepresentation: Johnny understands that $[_4$ I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$ $_4]$

If I am aware that Johnny got the message and my communicative intention has succeeded, I then entertain a fifth-order metarepresentation: I understand that $[_5$ Johnny understands that $[_4$ I intend that $[_3$ Johnny understands that $[_2$ I intend that $[_1$ he knows that $[_0$ he needs to post the letter $_0]$ $_1]$ $_2]$ $_3]$ $_4]$ $_5]$.

This is one of many characterisations of Gricean communicative intentions, so named given their origin with Grice (1957), who suggests that speakers' intention for their audiences to recognise their intentions is a central component of communicative meaning. The Gricean picture has been reformulated in a wide variety of guises (see e.g. Sperber and Wilson 1986; Moore 2014; Strawson 1964), but common to all Gricean accounts is the central role of speakers' intentions. Because it is difficult to accurately capture certain crucial features of human communication without appealing to speaker intentions, Gricean accounts persist – to the point of researchers writing of "exorcising Grice's ghost" (Townsend et al. 2017) from communication research, and despite some repeated and important critiques. One of these critiques is the accusation of cognitive implausibility: that is, Gricean communicative intentions, requiring as they do processing

of other individuals' intentions and other mental states with every utterance, cannot possibly be used constantly in communication, including by small children. The next section discusses this critique.

8.1.1 The plausibility of communicative intentions

The concern that metarepresentational abilities like these just seem too difficult – for adults, but especially for infants – has been widely discussed (Moore 2014; Geurts 2019; Moore 2016a). Because Gricean communicative intentions are taken to underlie every communicative interaction (no matter how simple that interaction is), engaging in these communicative interactions presupposes the presence of sophisticated social cognition – including the attribution of beliefs, inferences about the intentions of an interlocutor, and the ability to embed these concepts recursively (Moore 2016a).

Take the example of my one-year-old niece Lizzie raised in Chapter 2. In this interaction, I have noticed that there are balloons tied to the fence across the street and am attempting to draw Lizzie's attention to the balloons. Lizzie is occupied playing with her toys, so I am trying attract her attention by calling her name and pointing, saying "Lizzie! Look at the balloons!" My intention to draw Lizzie's attention to the balloons involves a first-order metarepresention on my part:

- I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence. $_0$] $_1$]] – a metarepresentation; that is, my intention to alter Lizzie's mental state by drawing her attention to the balloons.

If one of the balloons pops, making a noise and attracting Lizzie's attention, she will have seen the balloons, but my informative intention will not have succeeded, because Lizzie did not recognise it. This is one of the central insights of Grice, reformulated by Strawson (1964) and reiterated by Clark and Brennan (1991) and Tomasello (2008): that a successful informative intention entails the hearer's recognition of the speaker's intent to inform, as well as the content of the informative intention. Lizzie's recognition of my informative intention would require her to entertain a second-order metarepresentation:

- Lizzie is aware that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence $_0$] $_1$] $_2$]

Finally, my awareness of the success of my informative intention requires a third-order metarepresentation:

- I am aware that [$_3$ Lizzie is (/is not) aware that that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence$_0$] $_1$] $_2$] $_3$]

On the canonical account of ostensive-inferential communication, an informative intention alone is insufficient: the interaction requires a further level of metarepresentation involved in a communicative intention (Sperber 2000a; Sperber and Origgi 2012; Scott-Phillips 2015). That is, what would allow Lizzie to understand my intention to draw her attention to the balloons? To grasp this, she must understand why I am using the words and gestures that I am. At one year old, she is unlikely to have grasped the conventions behind these behaviours; she must therefore also understand *that the behaviour itself is communicative*, in the same way that Johnny must understand that my positioning of the letter on his desk is communicative. This requires Lizzie to entertain a fourth-order metarepresentation – my communicative intention:

- Lizzie is aware that [$_4$ I intend that [$_3$ Lizzie is aware that [$_2$ I intend that [$_1$ Lizzie perceives that [$_0$ there are balloons on the fence $_0$] $_1$] $_2$] $_3$] $_4$]

If the perception of communicative intentions is indeed necessary for language learning, infants would need to entertain at least fourth-order representations before beginning to acquire vocabulary. This commitment underlies the position that communicative intentions are what allow infants to acquire language (Tomasello 2008; Scott-Phillips 2015; Bloom and German 2000; Csibra and Gergely 2009; Csibra 2010). Even reducing the level of complexity to exclude the need for communicative intentions, and to require only the informative intention, leaves a one-year-old processing a second-order metarepresentation involving my intention and her own mental state.

Attributing a fourth-order, or even second-order metarepresentation to a one-year-old child intuitively seems like a stretch, but this concern is not based only on intuition: as discussed in Chapter 2 there is currently no uncontroversial evidence that infants of this age are capable of mindreading, and certainly no evidence that they are capable of recursive mindreading. The fact that young children engage in behaviours that appear to require communicative intentions is not evidence that this particular analysis of communicative intentions is an accurate description of the cognitive abilities involved.

Older children and adults are capable of recursive mindreading (Helming et al. 2015; O'Grady et al. 2015), but it is not clear that this behaviour is directly comparable to ostension, for two reasons. First, this evidence entails individuals being introduced to a single metarepresentional proposition over the course of a story, and being tested on parts of this proposition separately

in each question, whereas each act of ostension (that is, each utterance) requires rapid comprehension of a new metarepresentational proposition in its entirety. Secondly, the presence of this ability in an experimental setting does not necessarily imply that people deploy it constantly in communication; it may require more effort than is feasible for the kind of constant use that ostensive-inferential communication demands. Again, the continuous and effortless processing of communicative intentions by adults is not evidence that this particular analysis of communicative intentions is an accurate description of the cognitive abilities involved in this processing.

In summary, ostensive-inferential communication requires advanced mindreading abilities that infants may not develop early enough to use in language acquisition. It is also not clear that adults are capable of using this kind of advanced mindreading rapidly and flexibly enough for it to feasibly underlie everyday conversation (Moore 2016a; Geurts 2019). Because of this, a range of accounts have attempted to present an alternative account of communicative intentions that accepts the fundamental insights of Griceanism, but reduces the complexity of the mindreading required.

8.1.2 Minimalist alternatives

Not all proponents of Gricean accounts accept that the appearance of implausibility is weighty enough to lend credence to this critique. Scott-Phillips (2015) argues that the appearance of difficulty and complexity in formal descriptions of communicative intentions is not necessarily reflected in psychological reality. As an analogy, consider the mathematical description of the trajectory and velocity of a speeding car. This description would be complex: most people would have difficulty comprehending it, and those who could comprehend it easily would do so only with considerable training. Yet we calculate trajectories such as these without trouble when establishing whether it is safe to cross the street. Analogously, a complex formal description of communicative intentions does not *prima facie* mean that they are cognitively implausible.

It is true that people appear capable of many behaviours that require complex formal descriptions – folk physics, acquisition and use of complex morphology and syntax, singing, and so on. An account of the human capacity to judge the movement of physical bodies, create a syntactically well-formed utterance, or vocally generate a series of sounds at specific pitches nonetheless requires a proper description of how this is achieved in light of what is known about human cognition. In other words, the concern about plausibility could be rephrased as the question of how ostension is achieved in light of what is currently known about human cognition. I therefore

think the plausibility critique is worth taking seriously.

One response to the plausibility critique is to reduce the cognitive difficulty of the formal description by arguing for an easier class of metarepresentation: for instance, Breheny (2006) argues that ostensive-inferential communication can be achieved by representing less fully-fledged and demanding mental states. While representing a belief requires representing that another person holds a representational mental state that may be false (Apperly and Butterfill 2009; Martin and Santos 2016), Breheny (2006) argues that it may suffice to instead represent a mental state that is as simple as awareness or lack of awareness. This is a kind of mental analogue to level-1 perspective-taking – the ability to determine whether another agent can or cannot see something, without any additional information about their perspective (such as whether the angle they see it from will result in them seeing different parts of it) (Apperly 2011).

On this account, in my interaction with Lizzie, it would not be necessary that I intend that Lizzie *believe* that there are balloons on the fence; it is sufficient that I intend for her to *be aware of* the balloons. When Lizzie interprets my intentions, she does not need to represent my intention that she believe anything; she needs only represent my intention that she *sees* the balloons. This less demanding mindreading ability does not require being able to represent false beliefs, or beliefs of any kind; only a binary "being aware of *x*" or "not being aware of *x*." However, even if the mental states making up a communicative intention are less demanding than "belief", the communicative intention still requires recursive embedding of these mental states, and as such it is not clear that it fully solves the problem that metarepresentation seems intuitively difficult, regardless of the degree of difficulty of the representations involved.

Moore (2014, 2016) responds to the plausibility problem by developing a "minimally Gricean" account which, he argues, satisfies the main points of Grice's analysis, but does not require such advanced metarepresentational abilities. He argues that the Gricean account breaks down neatly into two separate actions: a sign, such as a point or an utterance, which encodes the speaker's message ("look at that") and is intended to invoke a behavioural response in the hearer; and an act of ostension, which directs the hearer's attention to the sign. This directing of attention to the sign, Moore argues, indicates to the hearer that the act is being performed communicatively for them, fulfilling the Gricean criterion that the hearer recognises that the sign is being produced as an informative intention (Moore 2014; Moore 2016a).

For instance, imagine that I am reading the newspaper over breakfast, while my flatmate Meredith has lost her keys and is in a hurry to leave the house. I saw the keys on the table before I opened my newspaper, which covered them. As Meredith passes me, I do not stop reading,

but pick up the keys from beneath my newspaper and put them down on the edge of the table. I have an informative intention: that Meredith knows where her keys are. If I simply move the keys without her noticing that I have done so, Meredith may perceive the content of my informative intention but not the intention itself – a failed informative intention – and so I allow the keys to make a small noise as I put them down on the table when she passes, drawing Meredith's attention to the fact that my moving them is an intentional, communicative act. Meredith's recognition of my making a noise as a communicative gesture that draws her attention to my informative intention creates a fully-fledged communicative intention, with all of the attendant levels of metarepresentation involved in my tracking of my own communicative intention. On a "maximalist" account, this would entail:

- I understand that [$_5$ Meredith understands that [$_4$ I intend that [$_3$ Meredith understands that [$_2$ I intend that [$_1$ she knows that [$_0$ her keys are on the table $_0$] $_1$] $_2$] $_3$] $_4$] $_5$].

On Moore's (2014, 2016) account, the movement of the keys is the sign: my message is "here are your keys". The timing and attention-getting noise of the sign, which direct Meredith's attention to the sign and make it clear that the sign is directed at her, make up the act of ostension (or act of address). This account, according to Moore (2014, 2016), substantially reduces the metarepresentational demands of a Gricean intention by breaking it into two much simpler, concurrent actions.

The sign, on this account consists of just a representation (Moore 2014): [$_0$ The keys are on the table $_0$].

Producing the act of ostension requires a first-order metarepresentation: I intend that [$_1$ You believe that [$_0$ my moving of the keys was an act produced for you $_0$] $_1$].

Meredith's comprehension of the act of ostension entails a second-order metarepresentation: [$_2$ She intends that [$_1$ I believe that [$_0$ her key-moving was aimed at me $_0$] $_1$] $_2$].

For a simple directive interaction, like Mary calling Andrea's name and then pointing to direct Andrea's attention to a cute dog across the street, the orders of metarepresentation are even simpler. Andrea's comprehension of the act of ostension requires a first-order metarepresentation ([$_1$ Mary intends that [$_0$ I hear and respond to her calling my name $_0$] $_1$]), as does her comprehension of the sign ([$_1$ Mary intends that [$_0$ I look at the dog $_0$] $_1$]) (Moore 2016a).

Moore (2014) notes a further advantage to his account, which is that the act of ostension may be part of a body of conventional and well-rehearsed physical actions, such as eye gaze and pointing (or, one could assume, language). These conventional ostensive acts would be so common that, most of the time, it would not be necessary to represent them (Moore 2014).

This account appears to successfully reduce the metarepresentational demands of a communicative intention. However, it has several important flaws which mean that it does not provide the simplification that it claims, and does not fully capture the phenomena involved in fully-fledged communicative intentions.

8.1.3 The gaps in minimalism

It is not clear that Moore's account achieves as much simplification as it appears to, because it requires interlocutors to comprehend two propositions concurrently. In the key-moving example above, for instance, Meredith must simultaneously entertain both "The keys are here" and "She intends that I believe that her key-moving was aimed me," which is a second-order metarepresentation.

To revisit our earlier example, with the original description of the informative intention, Lizzie must entertain "My aunt intends that I look at the balloons" which is a first-order metarepresentation, along with "My aunt intends that I hear and respond to her calling my name" – another first-order metarepresentation. While two concurrent first-order metarepresentations could plausibly be easier for a one-year-old child to comprehend than a single third-order metarepresentation, it is not self-evident that this is the case, and whether it is in fact more achievable for an infant is questionable.

More importantly, this account achieves its simplicity at the cost of failing to explain two of the primary phenomena that prompt such a complex metarepresentational description in the first place: the speaker's confirmation of the success of their informative intention, and the understanding of novel signals.

How do I, as the key-mover, know that my Gricean intention has succeeded? I moved the keys as Meredith passed, but perhaps she was distracted; maybe she was looking at her phone as she walked, glanced over at the table and saw her keys, and moved away. My informative intention has then failed; Meredith got the sign, but did not know that I had produced it for her.

However, if all I entertain is "your keys are here" and "you believe that my action was produced for you", as Moore suggests, I do not have any way of knowing that my intention has failed. In order to check this, I must observe Meredith's behaviour and see whether she noticed my act of ostension: I must wonder whether "Meredith perceives that I intend that she believed that my key-moving was aimed at her," – a third-order metarepresentation. Sensitivity to the success of an informative intention is something observed in infants as young as 12 months (Grosse et al. 2010; Liszkowski et al. 2008), but Moore's account achieves its simplification at the cost of

erasing this part of the interaction. If we include it, we are left in exactly the same place as the original account: the speaker must entertain a third-order metarepresentation, and the recipient must entertain a second-order metarepresentation – as well as both entertaining a separate representation: "the keys are here".

This account also gets us nowhere with the inference of meaning from novel signals that do not have a transparent meaning. It cannot explain Johnny's easy interpretation of why I left an envelope on his keyboard; laboratory results suggesting the ready recognition of communicative behaviours in minimal signals (Scott-Phillips et al. 2009); the emergence of new languages like Nicaraguan Sign Language; or the language-learning capabilities of infants who have had scant time to absorb conventional ostensive signals and understand that they are communicative.

I propose an alternative account drawing on the concepts of joint attention (Siposova and Carpenter 2019) and common ground (Clark and Brennan 1991). The following section gives an outline of the way I will use these concepts; demonstrates how communicative intentions can be described as instances of common ground; and details how this description can account for a wide range of communicative contexts.

8.2 Common ground: an alternative

8.2.1 Joint attention to shared attention

Let us begin with the commonly used concept of *joint* attention: the state of two individuals mutually attending to something, and being aware that they are mutually attending to that thing. When Lizzie and I look at the balloons together, we might look at each other, smile, and look back at the balloons, possibly looking back and forth or adding in other gestures like pointing or naming the colours of the balloons. We are both aware that we are engaged in attention to the balloons together.

Two people attending to the same thing are not engaging in joint attention unless they have an awareness that their attention is joint (Eilan 2005). If I am watching a dog at the park, while someone on a bench nearby watches the same dog, our attention is not joint until it becomes so by us becoming aware of this jointness, by making eye contact with each other and looking back and forth between each other and the dog. Similarly, if I see that someone on the next park bench is watching something, see that it is a dog, and begin watching the dog too, we are not engaged in joint attention unless we both recognise that we are both watching the dog (Eilan 2005).

Use of the term "joint attention" is somewhat inconsistent, sometimes being used to mean simply looking at the focus of another agent's gaze (e.g. Butterworth 1995) rather than this state of joint awareness. There is also a range of related terms, such as common ground (Clark and Brennan 1991), mutual manifestness (Sperber and Wilson 1986) and mutual knowledge (Wilby 2010). Siposova and Carpenter (2019) describe a typology of the range of concepts within the umbrella concept of "social attention", which they take to include any behaviour involving a triadic relationship between two interacting agents and an object. These "levels" of social attention differ on various dimensions such as the certainty that the agents feel in their jointness, their degree of shared background knowledge and closeness, and their goals. While joint attention is often described as a state of two individuals in a visual triadic relationship, social attention can apply to any type of sensory input (that is, people can listen together or taste together); objects of the attention can include past, future, imaginary, and non-existent states and events, including mental states; and multiple individuals can engage in a state of social attention (Siposova and Carpenter 2019).

Siposova and Carpenter (2019) delineate four separate stages of social attention, each distinct in its level of engagement and "togetherness". *Monitoring attention*, the lowest stage, is essentially sustained perspective-taking. For instance, if I notice that Lizzie is looking out of the window, I might follow her gaze and see that mounted police officers are passing on horses. I am aware of the object of her attention (the horses) and her attention to the object, but she is not aware of my attention. I am monitoring her attention, but there is no jointness. In Stage 2, *common attention*, two individuals may each be aware of each other's attention and aware that they are both attending to the same thing, but the experience is still fundamentally individual. I may be in common attention with the person next to me at the cinema; they are facing the screen with their eyes open, and so I assume that they are watching the film, while they probably assume the same about me (and I could surmise that they probably assume it about me, etc.), but we are not yet engaged in joint perception (Siposova and Carpenter 2019).

It is only with *mutual attention*, Stage 3, that attention becomes truly joint, in that there is bidirectional contact between two individuals. This can occur with as little as a moment of brief eye contact, which allows reciprocal confirmation of the attention of each agent and (with facial expressions) even some reactions to the object. If Lizzie turns around to see what I am looking at, and sees my gaze move from the horses to her, we will enter a state of mutual attention to the horses outside the window. With intentional communication about the object of attention comes Stage 4, *shared attention* (Siposova and Carpenter 2019). It is this state of shared attention

that can be used to characterise Gricean communication.

In the introduction to this chapter, I used the term "joint attention", since this is a more recognisable key phrase. Having introduced this framework, I will use the terminology introduced by Siposova and Carpenter (2019) – that is, I will use the term "mutual attention" to refer to reciprocal awareness of attention between two individuals, and "shared attention" to refer to intentional, synchronous communication about a subject of mutual attention. Shared attention and mutual attention should be understood as types of joint attention in its strictly defined sense; I will use these terms exactly as defined above, in order to avoid the ambiguity that may arise with the more widely and imprecisely used term "joint attention". When discussing other arguments or accounts, I will use the terminology used in those arguments.

8.2.2 Recharacterising communicative intentions

In a state of shared attention, individuals can direct each other's attention, jointly focus on an object, and confirm the other's perception of that object. Pointing to direct an interlocutor's attention to an object is done with the goal of making that item the object of shared attention ("Look at the balloons, Lizzie!"). A Gricean informative intention, in these terms, has the goal of bringing the content of the message into shared attention.

When I move the keys to the end of the table for Meredith to see and want confirmation that she has recognised my informative intention, I will seek this confirmation by looking at her and entering into a moment of shared attention with her in which we are both attending to the keys, and aware that we are both attending to the keys. Her mental state (and mine) could be described as "We are now both aware that the keys are here." The success or failure of my informative intention rests on whether this attempt to induce shared attention is achieved: if she enters into this moment of shared attention, it has succeeded, but if she was distracted during my action, spots her keys afterwards and takes them without engaging in shared attention, my intention has failed. Importantly, "We know that the keys are here," is not metarepresentational: it is a shared first-order representation, and becomes metarepresentational only if it is introspected, as in "We/I know that we know that the keys are here." I will return to what it means for a representation to be "shared" in Section 8.3.2.

A similar state can be achieved *conceptually* rather than *physically*. For instance, when I tell my friend that I saw Robert yesterday, we are in a state of shared attention to the concept of Robert, even without Robert's physical presence. In this way, shared attention may apply to physical occurrences or entities that are temporally or spatially removed (if I tell my friend that there

was a thunderstorm in London today, we jointly attend to the idea), or even to abstract concepts like love (Heal 2005; Tomasello 2018; Siposova and Carpenter 2019). Eilan (2005) argues that "words are, on this conception, an immensely delicate and useful way of pointing."

These examples all use declarative utterances, but the desire to bring about a state of shared attention can apply to any speech act. For instance, when I ask, "Is it raining out?" my assumption is that my interlocutor has information I do not have, and my intention is to bring my interlocutor's information into shared attention. When I say, "I promise that I'll respond to your email before the end of the week," I am creating a state of shared attention to the email and my plans for it. "I now declare you married," creates a (group) state of shared attention to cultural concepts of marriage and its significance for the couple being married.

Shared attention can only account for certain types of communication, because it is essentially a synchronous experience. Asynchronous communication, such as writing letters, does not allow a true, real-time experience of shared attention to the topic of conversation (Siposova and Carpenter 2019). Here, the concept of common ground may help us. This is the the information that interlocutors are aware of sharing (Clark and Brennan 1991), and it does not require the synchrony of shared attention. Where shared attention requires both individuals to be attending to the same concept simultaneously, and to be aware of the jointness of their attention at the same time, common ground allows for asynchrony. Both individuals must be aware that the information in common ground is shared, but do not necessarily need to be attending to it simultaneously for it to be in common ground. Given the vast content of what is in common ground, from shared cultural context to the entirety of two people's shared experiences (Clark and Brennan 1991), it would not be possible for everything in common ground to be attended to simultaneously.

Extending the account of Gricean intentions to include common ground broadens the type of communication that can be accounted for: that is, the goal of all informative intentions is to move certain information into common ground. In certain cases, the goal is additionally to achieve a state of shared attention. When I send a text to my father telling him that I am home safe and sound, I do not know whether my informative intention has been successful until my text is marked as having been read. Once the text is marked as read, even without a reply, my father and I can both be confident that we now share this information – we are both aware that I am home. In the case of Johnny and the letter, my confirmation of the success of my informative intention is delayed until I know that Johnny has understood my signal, perhaps when I get home and he tells me that he decoded the message. At that point, we are both aware that the need

for him to have posted the letter is shared information. An asynchronously delivered message (such as an email) requires confirmation that the informative intention has been received just as much as an attempt to instigate shared attention: if the email does not receive a reply within the expected timeframe, it may prompt a second message to ensure that the informative intention has succeeded.

Shared attention cannot be the goal for asynchronous communication, but may also not be the goal even in complex synchronous communication. For instance, when a friend explains to me how to make fresh pasta, her explanation will likely be complex enough that we will not be jointly attending to exactly the same concepts in any given moment. We are in a state of shared attention to the general topic, but likely not to the moment-by-moment concepts. Nonetheless, throughout the interaction, my minimal responses assure her that her explanations are being understood, and the information is becoming part of our common ground (Clark and Brennan 1991). Feedback from the receiver of an utterance – either repair that clarifies the intended meaning of the utterance, or confirmation that the receiver believes she has understood the utterance – gives the signaller evidence that the content of their signal has entered common ground, and their informative intention has been successful (Clark and Brennan 1991). I will use "common ground" as an umbrella concept that includes instances of shared attention, which are a specific instance of common ground where the information that interlocutors share includes their state of shared attention.

Common ground can thus provide an alternative description of the third-order metarepresentation of a Gricean informative intention – the content of the message, the receiver's awareness that the signaller intends to inform them of this content, and the signaller's awareness that the receiver has recognised the informative intention. In the example with Meredith and the keys, the third-order metarepresentation that is my successful informative intention can be reformulated. The original formulation:

- I understand that $[_3$ Meredith understands that $[_2$ I intend that $[_1$ she knows that $[_0$ her keys are on the table $_0] _1] _2] _3]$.

now becomes:

- We know that $[_0$ her keys are on the table $_0]$.

By reframing informative intentions as instances of shared attention, this account reformulates a third-order metarepresentation as a shared representation. An informative intention is

an intention to change a proposition (X) from being unshared information into information that is in common ground. X could be anything – a request to post a letter, the location of keys, the presence of balloons, the warning "there are baboons in the road", the recommendation to take an umbrella, or an understanding of how to make pasta from scratch.

This is not yet a full account of communicative intentions, since ostensive-inferential communication also requires an explanation of how an audience recognises that a speaker has an informative intention. How does Meredith know that I am attempting to induce a state of shared attention to the keys – how does Meredith know that I intend that $[_1$ we know that $[_0$ her keys are on the table $_0]$?

8.2.3 Ostensive behaviour

In the characterisation of communicative intentions described in Chapter 2 (Scott-Phillips 2015; Sperber 2000a), the inference that a speaker is trying to communicate is explained with additional layers of metarepresentation. That is, Meredith understands my informative intention:

- Meredith understands that $[_2$ I intend that $[_1$ she knows that $[_0$ her keys are on the table $_0] _1] _2]$.

How does Meredith understand that I am trying to inform her of something? According to Scott-Phillips (2015), she understands this because I move the keys *ostensively*, perhaps in an exaggerated manner, or with other behaviour that informs her that my behaviour is communicative. The point of my communicative intention is to make Meredith aware of my informative intention:

- I intend that $[_3$ Meredith understands that $[_2$ I intend that $[_1$ she knows that $[_0$ her keys are on the table $_0] _1] _2] _3]$.

Meredith's comprehension of this communicative intention is what allows her to interpret my behaviour as communicative, and use relevant contextual information to infer the content of that informative intention. Meredith's comprehension of this, and my recognition of her comprehension of it, is what leads to the fifth-order metarepresentation outlined above:

- I understand that $[_5$ Meredith understands that $[_4$ I intend that $[_3$ Meredith understands that $[_2$ I intend that $[_1$ she knows that $[_0$ her keys are on the table $_0] _1] _2] _3] _4] _5]$.

If the informative intention is reframed as an instance of shared attention, this becomes:

- I understand that [$_3$ Meredith understands that [$_2$ I intend that [$_1$ we know that [$_0$ her keys are on the table $_0$] $_1$] $_2$] $_3$].

However, spelling out the metarepresentational scenario does not explain *how Meredith recognises this communicative intention*. Ending the explanation here begs the question: Meredith recognises that I have an informative intention because my behaviour is communicative. How does she know my behaviour is communicative? She recognises my communicative intention. How does she recognise my communicate intention? My behaviour is ostensive. What is ostensive behaviour? It is behaviour that expresses a communicative intention.

The metarepresentational chain itself does not explain how individuals recognise communicative intentions (note that this does not imply that the individuals involved cannot have metarepresentations of the informative intention, its recognition, and so on – see Section 8.3.4 for further discussion of this). The "exaggerated behaviour" mentioned by Scott-Phillips (2015, p. 9) comes closer to being an explanation of how ostensive behaviour is recognised, but this can only explain a very limited range of ostensive actions, like chewing ostensively ("I can't respond yet because my mouth is full"). It cannot explain the recognition of an eyebrow lift, the placement of a letter, or swiftly moving an on-screen avatar back and forth as ostensive.

A fuller explanation of how ostension is recognised can be established by contrasting a scenario in which I want my communicative intention recognised with one in which I would prefer to keep them hidden. This is the concept of *hidden authorship* (Tomasello 2008): the intentional disguise of informative intentions. Imagine that Meredith has been looking for her keys for ten minutes before I realise that I accidentally covered them with my newspaper. Embarrassed, I attempt to avoid drawing her attention to the fact that I have made her late, and so I quietly slide the keys out from under my newspaper and to the edge of the table when she is not looking, ready for her to spot them when she next walks past the table. My goal is still for Meredith to see her keys, but it is not to bring about a state of shared attention to the position of the keys.

The difference between the two scenarios is similar to Moore's "act of ostension": in the informative intention scenario, I make a noise with the keys, time the movement of them to coincide with Meredith's presence, and meet her eye when her attention is grabbed; in the hidden authorship scenario, I do nothing to attract or join her attention. Attracting attention is a necessary characteristic of ostension – it is not possible to move information into common ground with someone without first attracting their attention – but is not sufficient. When someone sneezes loudly on a train, they attract my attention, but I do not assume that they are trying to communicate something to me. In order to be ostensive (that is, to signal its own signalhood), a behaviour

needs to be not just attention-attracting, but also to have no reasonable alternative explanation. This is an extension of the "act of ostension" of Moore's account, which requires just attention attraction (Moore 2016a; Moore 2014).

The lack of a reasonable alternative explanation for an attention-grabbing behaviour can explain the inference of an attempt to communicate in a wide range of communicative contexts:

- When Johnny comes home to find a letter propped up vertically on his keyboard – clearly put there by me, since no-one else has access to the flat – he will realise that this is odd behaviour on my part; there is no reason for me to put the letter in such a place unless I am trying to communicate something. In trying to ascertain what I was attempting to communicate, he will draw on our shared context (the urgency of posting the letter, the fact that I was meant to have posted it, and the fact that we have both been out of contact all day) to infer what I was trying to communicate.
- When an experiment participant moves an on-screen avatar frantically back and forth between two squares on a grid in a communicative game where this movement serves no other purpose and there is no dedicated communication channel, their partner in the game can assume that the frantic back-and-forth movement is communicative, and begin attempting to infer the meaning (no mean feat in itself, but a separate task from establishing that the behaviour is communicative) (Scott-Phillips et al. 2009).
- If the car in the valley below me, a few cars ahead of me, flashes its hazard lights for a few seconds; and then the next car in the queue flashes its hazards; and then in the car in front of me repeats this behaviour, the standard reason for hazard lights – for example, "Be cautious; I need to stop unexpectedly" – will not explain this behaviour. Instead, I must find some other reason for the unexpected behaviour, such as warning of an obstacle in the road (or, in this real-world case, a troop of baboons).
- When someone points at something, calls my name, or tells me that it's raining out and I'd better take an umbrella, they are using conventional communicative channels, for which there is no explanation other than their goal of communication.
- If I'm heading out the door and my partner calls "It's raining out", his desire to make a banal comment about the weather is technically possible, but implausible; his behaviour (his utterance) has no reasonable explanation other than an intention to communicate something unsaid. In this case, his goal is to give me a message beyond what is encoded by his words, such as "You might want to take an umbrella."

On this account, ostensive behaviour (attention-grabbing behaviour with no plausible non-

communicative explanation) clues a receiver in to the fact that a signaller is attempting to communicate something. This means that the additional layers of metarepresentation are not required to explain how an informative intention is recognised: informative intentions are recognised because they are the only explanation for otherwise inexplicable behaviour. To return to Meredith and the keys, my informative intention is:

- I intend that [$_1$ we know that [$_0$ her keys are on the table $_0$].

Meredith recognises my informative intention because there is no other explanation for my exaggerated, noisy way of moving the keys, accompanied by eye contact – I can move keys perfectly well without clattering them, and I do not need to move the keys or make eye contact with her in order to have breakfast and read the newspaper.

To summarise, any attention-seeking behaviour that does not have a reasonable alternative explanation is understood as ostensive behaviour. Ostensive behaviour can be non-linguistic, as in the examples of moving keys noisily, putting a letter in an odd place, or chewing in an exaggerated way. It can be non-embodied, as in the case of an experimental game that requires participants to communicate exclusively by moving avatars on a grid, or in the case of using signals on a car to communicate novel messages. Conventional communicative behaviour (pantomime, language, pointing) is, by definition, ostensive, since there is no alternative explanation for it. This definition of ostension applies at the level of the utterance: there is no explanation for me saying "It's raining" in response to "I'm going out" other than me having an informative intention. Ostensive behaviour prompts an audience to recognise that a signaller has an informative intention. The goal of an informative intention is to create a shared representation by moving information into common ground.

The common ground account bears a great deal of similarity to a range of different discussions of Gricean communication, shared attention and common ground. The next section discusses some of these similarities, before I turn to discussing a range of problems and questions raised by the common ground account in Section 8.3.

8.2.4 Finding common ground in common ground

As mentioned above, Eilan (2005) draws a comparison between language and pointing (arguably the prototypical gesture for instigating shared attention), suggesting that language in general fulfils, in a much more complex manner, essentially the same function as basic instances of shared attention. Siposova and Carpenter (2019) note that conversational partners share knowledge

about the topic of their communication, as well as the act of communication itself (that is, the fact that they are intentionally communicating about the topic). Tomasello (2008, p. 91) explicitly connects Griceanism with both shared attention and common ground:

> *"Overt expression of the Gricean communicative intention places the communicative act itself – the gesture or the utterance – into the participants' common ground, specifically, into the ongoing shared attentional frame within which they are communicating. Thus, it is most precise to say not just that I want you to know that I want you to attend to something, but that I want us to know this together."*

Peacocke (2005) suggests that in extended instances of joint attention, the participants create a "mutual world [involving] the events which are jointly attended to, the participants' relations to them, the development of both of these over time, and what is known at each stage about what has happened earlier in the mutual world of joint attention."

A similar idea is also recognisable in the concept of "mutual manifestness" (Sperber and Wilson 1986) – the features of an environment that are readily perceptually accessible to anyone in that environment, including the feature that everyone shares the same cognitive environment. Sperber and Wilson (1986) define manifestness is a state of perception, awareness, or inference that is "weaker" than knowledge. That is, while knowing a fact entails having a mental representation of it, a fact that is manifest may be "dormant", like the manifest assumption that Nelson Mandela never walked on the moon. This means that mutual manifestness is "weaker", in the same sense, than mutual knowledge. Wilson (2013, p. 135) argue that this makes mutual manifestness "more realistic, more psychologically relevant, and at least as cogent as the notions of mutual knowledge, common knowledge, or common ground."

On this account, ostensive behaviour is an attempt to make features that are manifest, or salient, to one individual equally salient to their interlocutor, so that they become mutually manifest. Realistically, there are countless phenomena within a cognitive environment that are manifest, and that a communicator may be attempting to make mutually manifest. Sperber and Wilson (1986) suggest that it is the *principle of relevance* that allows audiences to determine which of these features the speaker is attempting to make manifest. That is, ostensive communication automatically carries a guarantee that the information the speaker is trying to communicate will be relevant to both speaker and audience, allowing the audience to easily settle their attention on the relevant features of the environment, now mutually manifest to speaker and audience.

These accounts of human communication have important differences, as made clear in the distinction between mutual manifestness, mutual knowledge and common ground Sperber and Wilson (1986). A full description of these accounts and the distinctions between them is beyond the scope of this chapter, but in common is the emphasis on *jointness* in communicative intentions, and the recognition that ostensive communication is an attempt to create a symmetrical and shared state of awareness of the content of the communication.

The reformulation of communicative intentions as essentially joint has two crucial advantages over other characterisations. First, it opens the door to the infant joint attention literature. Drawing on the developmental shared attention literature may help to unpick the developmental knot that requires mindreading for language acquisition, but language acquisition for mindreading (see Chapter 3). This is because, unlike mindreading, joint attention uncontroversially emerges early in human development. Heal (2005, p. 5) gives the following account of its developmental trajectory:

> *"In the very early weeks and months of life we find carer and child attending intensively to each other – gazing into each other's eyes, smiling at each other, copying gestures, and the like. From about 6 months we note the appearance of outward-directed attention, occurring in situations where it is natural to think of three corners of a triangle: namely, carer, infant, and a part of the world to which both attend. First to appear is a tendency in the infant to follow the outward-directed gaze of the adult. Then there is development of more sophisticated variants, such as gaze-checking if the initial gaze-following does not easily identify something worth attending to. Bit by bit, yet more elaborations follow, such as attempts by the child to bring about joint attention by use of pointing gestures."*

By approximately 12 months, infants engage in sustained bouts of triadic joint attention – attending to an external object along with someone else (Eilan 2005). If early experience with shared attention is sufficient to enable the acquisition of initial vocabulary, and in turn develop richer communicative behaviours that bootstrap both language acquisition and the development of the skills that ultimately result in success on explicit false belief tasks, then shared attention can explain how young children can be ostensive communicators without relying on mindreading.

Note that this claim does not necessarily commit one to the position that young children *cannot* engage in any mindreading behaviour; rather, it avoids the need to rest explanations of infant language acquisition on the current uncertainties of the infant mindreading literature, and

suggests that the most fruitful empirical direction for understanding infant and adult ostensive behaviour may be shared attention rather than mindreading.

The second advantage of this account is that it offers a more complete and accurate psychological description of what is involved in ostensive communication. In order to understand why this is the case, we must turn to the problem of infinite regress in shared attention. Section 8.3 elaborates on the problem of the infinite regress, discusses a range of proposed solutions to it, and explains why these solutions mean that a common ground account is more complete than a metarepresentational one.

8.3 Sidestepping the infinite regress

Metarepresentational accounts of both ostensive-inferential communication and common ground both face the problem of dissolving into an infinite regress. Returning to Johnny and the letter, a metarepresentational account of my comprehension of his having understood my communicative intention requires a fifth-order metarepresentation:

- I understand that [$_5$ Johnny understands that [$_4$ I intend that [$_3$ Johnny understands that [$_2$ I intend that [$_1$ he knows that [$_0$ he needs to post the letter $_0$] $_1$] $_2$] $_3$] $_4$] $_5$].

How do I know my communicative intention has succeeded? Probably because Johnny has told me. Perhaps he texted me, or left a note on the fridge, or told me when I walk in the door. He is aware that I'll want to know that he got the message. In other words:

- Johnny intends that [$_6$ I understand that [$_5$ Johnny understands that [$_4$ I intend that [$_3$ Johnny understands that [$_2$ I intend that [$_1$ he knows that [$_0$ he needs to post the letter $_0$] $_1$] $_2$] $_3$] $_4$] $_5$] $_6$].

If I realise that Johnny got the message – perhaps I see the letter missing from his desk before I see the note on the fridge – and Johnny notices this, he is likely to draw my attention to his explicit message. So his intention to inform me is more than him simply wanting me to know that the letter has been posted; it is an informative intention, with the attendant levels of metarepresentation, and the content of which is my confirmation of the success of my communicative intention. That is:

- Johnny intends that [$_8$ I understand that [$_7$ Johnny intends that [$_6$ I understand that [$_5$ Johnny understands that [$_4$ I intend that [$_3$ Johnny understands that [$_2$ I intend that [$_1$ he knows that [$_0$ he needs to post the letter $_0$] $_1$] $_2$] $_3$] $_4$] $_5$] $_6$] $_7$] $_8$].

Of course, this informative intention is embedded inside its own communicative intention, taking it up to ten levels of metarepresentation. And perhaps there is a covert message here too – perhaps Johnny is not generally particularly reliable, and wants to make it clear that he is making an effort; he will want to know that I have seen and understood this effort. If I make it clear that his message has been received, we begin another round of embedding. Continuing to spell out the metarepresentations in this way therefore becomes an infinite regress.

8.3.1 The problem of coordinated attack

This problem is essentially one of "coordinated attack" (Fagin et al. 1995; Wilby 2010). To illustrate, imagine two military units on two hills, with a valley in between them in which the enemy waits. The two units must attack the enemy simultaneously in order to be successful; one unit on its own would be doomed to failure. The commanders of the two units must therefore attack simultaneously, or not at all. Commander 1 sends a messenger – the only means of communication – to the other camp suggesting a plan of attack, and waits for a reply. Commander 2 agrees with the plan, and returns the messenger with a note expressing assent. But now Commander 2 is uncertain: what if the messenger has been waylaid on the dangerous journey between camps? If Commander 1 has not received the message, Commander 2 is at risk of attacking alone. So Commander 2 must now wait for a message from Commander 1 that assent has been received, and the attack will go ahead as planned. As soon as Commander 1 sends the messenger with this agreement, the same uncertainty rears its head: if the messenger is waylaid, Commander 2 will not know that the plan is secure.

The central point here is that no number of messages will suffice for both parties to be absolutely sure that they share the plan of attack. This is the problem with Moore's (2014, 2016) minimalist account mentioned above: the more minimal demand of a speaker simply intending the speaker to comprehend "your keys are here" and "you believe that my action was produced for you" does not allow for speakers to confirm that their informative intention has succeeded.

Recall that this is a central feature of the Gricean account: that an informative intention entails me not just wanting Meredith to know where her keys are, but also wanting her to know that *I want her to know where her keys are*. If Meredith simply spots the keys where I have placed them without recognising that I placed them there intentionally, I would be very likely to make some other attempt to draw her attention to the fact that I wanted her to know where are keys are – perhaps remarking "Ah, good, you saw them then," or "Sorry, they were under my newspaper." Children as young as 12 months have been found to be sensitive to the success of their

informative intentions (Grosse et al. 2010; Liszkowski et al. 2008); it is an essential explanandum of an account of ostensive communication. This need to confirm the success of an informative and communicative intention, though, is what sets off the infinite regress, as demonstrated in how my message for Johnny could play out.

A satisfactory explanation of common ground faces the same problem. For instance, Battich and Geurts (2018, p. 3) characterise certain analyses of common ground (Lewis 1969; Schiffer 1972; Geurts 2018) as giving rise to infinitely recursive structures such as:

> *"p is common knowledge between A and B iff A knows that p, B knows that p, A knows that B knows that p, B knows that A knows that p, and so on ad infinitum."*

To reformulate Johnny's and my interaction in this way, *p* is the content of my initial informative intention: Johnny needs to post the letter. I know that *p*, Johnny knows that *p*, and Johnny knows that I know that *p*. Johnny's confirmation message means that I know that Johnny knows that *p*, Johnny knows that I know that Johnny knows that *p*, and so on.

Wilby (2010) addresses the regress problem in mutual knowledge, which is a recognisably similar concept to common ground: "knowledge of a situation, event or object [that] is out-in-the-open between two people" (Wilby 2010, p. 84). Wilby argues that the problem of coordinated attack illustrates that an infinite regress is an unavoidable feature of any account of mutual knowledge that rests on a back-and-forth account of individual mental states:

> *"The problem that this paradox uncovers is not just structurally similar to that of mutual knowledge, but simply* is *the problem of mutual knowledge illustrated within the context of coordinated action: it outlines a situation in which nothing less than mutual knowledge would suffice for appropriate action, and then raises a problem of how such mutual knowledge could be achieved."* – Wilby (2010, p. 87)

Quite apart from the psychological implausibility of an infinite regress, characterising common ground as a recursive sequence of states of knowledge fails to capture the state of "openness" that characterises the phenomenon. To illustrate, Wilby (2010) gives the example of two spies covertly watching each other look at a fireworks display. Each is aware of the other's awareness of the fireworks: Alex knows that Bruce perceives the fireworks, and Bruce knows that Alex perceives the fireworks. Then perhaps Alex realises that Bruce is actually watching him too: Alex now knows that Bruce knows that Alex perceives the fireworks. If Bruce has the realisation that Alex is watching him, then Bruce knows that Alex knows that Bruce perceives the fireworks. That is:

- A knows that B knows that A knows that *p*
- B knows that A knows that B knows that *p*

Despite this recursive sequence of states of knowledge, Alex does not yet know that Bruce has cottoned on to being watched; and Bruce does not yet know that Alex has cottoned on. There is not a state of shared attention, in which both spies are fully aware of their mutual awareness of each other and the fireworks. Each spy could individually keep adding layers to their individual awareness, without *p* ever entering a truly mutual state of common ground.

The infinite regress illustrates a profound problem with characterising a state of common ground as consisting of a recursive chain of individual mental states, since every stage of the regress fails to capture the openness of common ground. Clearly, a solution to the problem must be both psychologically plausible, given the central role that this phenomenon plays in social cognition; and it must properly capture the *jointness* of the phenomenon. In the next section, I describe various attempts to deal with the infinite regress, arguing that Wilby's (2010) argument offers the most convincing response.

8.3.2 Routes out of the regress

One suggested solution to the problem of regress is to move away from defining common ground in terms of component mental states like "knowledge" and instead rely on states like "perception" (e.g. Campbell 2005). That is, when Lizzie and I are jointly attending to the balloons, her shared attention to the balloons makes up part of my visual experience, and my shared attention makes up hers; it is a perceptual, rather than cognitive, state. However, this cannot account for the broad range of behaviours and contexts that make up ostensive-inferential communication. It fails immediately with any state of shared attention to an object distance in time or space, or to an abstract concept, since shared attention to a concept cannot be a perceptual state.

Battich and Geurts (2018) characterise the idea of the regress as arising from the misunderstanding that A and B must make an infinite series of inferences and represent the outcomes of these inferences in order to have common knowledge. Lewis (1969, p. 53) argues that "this is a chain of implications, not steps in anyone's actual reasoning. Therefore there is nothing improper about its infinite length." It may be true that these are not steps in anyone's actual reasoning, but that simply defers the problem. What *are* the actual steps in people's reasoning? What is a better psychological explanation of a state of common ground?

Tomasello (2008, p. 96) suggests that we compute the regress "as far as we need to or are able to, which is typically only a few levels up." This does not address the point that any given level of

the regress does not describe a state of true jointness, as illustrated by the problem of coordinated attack. It is the infinity *itself* that appears to provide the mutuality of common ground, meaning that a psychologically plausible account cannot simply opt out of addressing the regress (Wilby 2010).

Wilby (2010) argues that the way out of the regress is to understand common ground as something that cannot be reduced to the individual mental states of the participants. Instead, a "relational" analysis treats jointness as a psychological primitive: ascribing a state of joint attention to Lizzie implies that there is someone with whom she is co-attending. Lizzie and I do not make up components of each other's knowledge, but instead are joint subjects: *we* know that *p*. That is, Wilby (2010, p. 93) argues, "the participants are in direct, unmediated cognitive contact with each other to the extent that they literally share the mental state of mutual knowledge."

The advantage of this conceptualisation allows for mutual knowledge to play an explanatory role in joint attention in the same way that individual knowledge plays an explanatory role in individual attention: as an unanalysed and unanalysable primitive psychological state. Instead, participants' mental states are best characterised as "we know that Johnny needs to post the letter"; "we know that Johnny got my message"; "we know that Johnny is making an effort." Wilby (2010) acknowledges that this nonreductive concept may appear somewhat mysterious, and suggests that the best way of understanding it is to appeal to ontogeny. From their very earliest experiences, infants develop an understanding of themselves as part of an "us" from the very beginning of life, and as a central feature of human social cognition (Heal 2005; Wilby 2010).

This is plausible, but does not quite dispel the mystery about what it means that "we know *p*." I suggest that the essential difference of this second-person state is evident in the distinction between an instance of shared attention and the comprehension of a metarepresentational chain involving different people. For instance, take the condensed plot of Othello, as described in Chapter 2:

> "*[$_5$ Iago intends that [$_4$ Cassio believe that [$_3$ he intends that [$_2$ Desdemona intend that [$_1$ Othello consider [$_0$ Cassio's rehabilitation $_0$] $_1$] $_2$] $_3$] $_4$] $_5$]*." – Van Duijn (2010)

This is a difficult proposition to comprehend in its entirety. It may be easier to follow when established step-by-step in a play, but holding all of the different individuals' mental states in mind at the same time is challenging. By contrast, it is very easy to grasp the entirety of the meaning of this:

> "*[$_5$ I know that [$_4$ you know that [$_3$ I know that [$_2$ you know that [$_1$ I know [$_0$ who*

wrote Othello $_0]$ $_1]$ $_2]$ $_3]$ $_4]$ $_5]$."

That is – *we know who wrote Othello*. It is significant that one of these metarepresentations involves a group of individuals, while the other involves a back-and-forth between the same two individuals. The distinction is that mutual knowledge does not require either of us to represent the other's mental state as something separate and independent from our own; I can represent your mental state as simply being identical to mine. This can extend to more than one additional person. For instance, if I watch a film with three friends, I can include these three friends in my own states of awareness and knowledge: I know this, and these three people know it with me. That is, *we* know it, together.

By contrast, the behaviours that are generally identified as mindreading require considering differences in perspective. For instance, understanding how someone else's intentions explain and predict their behaviour, or determining what they do and do not know, or how an object appears to them, all demand that I consider another agent as a third person, rather than as something more like an extension of myself. The results of the explicit, single-avatar DPT may be considered in this light: the altercentric effect is a demonstration that a shared perspective with another agent (although this is not shared in the true sense of joint attention, since the other agent in this case is not real) is processed more rapidly than an unshared perspective.

Conceptualising communicative intentions in this way – as *I intend that we know A* rather than *I intend that you know that I intend that you know that I intend you to know A* – changes the questions we need to answer to understand the phenomenon. Most importantly: what allows me to safely believe that we are in a state of shared attention, what role does mindreading play in this assumption, and what role remains for mindreading in communicative intentions generally?

8.3.3 The role of mindreading

The common ground account of communicative intentions does not negate the role of mindreading in language. Rather, it constrains it and identifies more specifically the points at which it plays a role, which are numerous.

Beginning with ostensive behaviour, mindreading must play a role in determining that there is no other explanation for a given behaviour. In order to establish that I must have left a letter in an odd place in order to be communicative, Johnny must consider my motivations and intentions as separate from his own, and both imagine and discount other possible explanations for the behaviour. For my part, I must consider what Johnny will consider odd and therefore

communicative behaviour. Deciding how to behave ostensively, and interpreting ostensive behaviour, both therefore still require making assumptions and inferences about another individual's mental state – mindreading, albeit not metarepresentation. This raises the question of how infants interpret the kinds of communicative behaviours directed at them (more likely to be pointing and simple language than subtle facial expressions or the placement of objects). Proponents of "natural pedagogy" attribute it to a human-specific adaptation that predisposes infants to interpreting certain behaviours as ostensive (Csibra and Gergely 2009; Csibra and Gergely 2011); I will return to this in Section 8.4.

Intending to move information into common ground necessarily requires determining that said information is not already in common ground, which requires mindreading. Similarly, deciding how to package that information, and how much context to include, requires an understanding of the audience's existing context and knowledge. On the part of the audience, mindreading is necessary for inferring what information the signaller is attempting to move into common ground. This is the point at which the principle of relevance (Sperber and Wilson 1986) plays a powerful role, by constraining the range of inferences about a speaker's intentions to those that are relevant to speaker and hearer.

The inferences made at this stage may in some cases be very constrained and obvious, and in other cases be a substantial leap with little evidence. With young children, there is very little inference involved at all; children's communicative abilities are so limited that simply drawing shared attention to an object is not always successful. When I call Lizzie's name and point to the balloons, the content of my message ("The balloons are pretty") is very easily recoverable from my words and gestures.

In other cases, contextual information may play a significant role. In the case of Johnny and the letter, his inference depends on all of the background knowledge of what the letter is, why it is time-sensitive, the details of our arrangement, and even the knowledge that I have a habit of sending hurried messages by leaving items in odd locations. Given this common ground, the inference itself may not be particularly difficult for Johnny. In the reverse case, the example of an experimental communication game that provides no established communicative channel does not allow for a great deal of context sharing, making leaps of inference that much more difficult: if my partner is moving a figure frantically back and forth between two squares, I may easily infer that she is trying to send a message, but without much shared context, I will have to guess whether her message is "red", "green", or "blue". That is, a high degree of pre-established common ground may increase the likelihood of successfully achieving shared attention (Siposova and

Carpenter 2019).

In everyday language use, utterances may be underdetermined, but may nonetheless require more or less context to interpret. The utterance "Better take an umbrella – it's raining" does not require any detailed common ground specific to the interlocutors to understand. Although there may be some individual background knowledge involved (perhaps the speaker knows that the hearer particularly hates getting wet, and that their journey will not be of an insignificant distance), and there is some cultural common ground here (people generally prefer not to get wet), the utterance does not require a great deal of context to interpret. By contrast, "She left it over there" requires considerable context (and therefore understanding of what is in common ground) to interpret.

The extent of the inference needed to interpret a communicative intention may explain a failure to understand the message correctly. There may be a failure to converge on the most relevant information, as in the example in Chapter 2 – I intended to communicate to my friend with a covert glare that the speaker was a bore, but she, having tuned out of the conversation, assumed that different contextual information was the most salient, and interpreted my glare to be a complaint about the music. There may also be a failure to correctly infer a speaker's intentions: perhaps Johnny could infer that I changed my mind about something in the letter that needed posting, and left it on his keyboard merely to inform him that I didn't send it after all. Greater saliency of an object of attention may be more likely to bring about a successful state of shared attention; a moment of shared eye contact with a stranger may have a very obvious meaning if a loud alarm has gone off, but less obvious if the shared environmental object is more subtle, such as an annoying but quiet beeping from the traffic light outside (Siposova and Carpenter 2019).

This raises an important point: communicative acts may increase in their requirements for both contextual knowledge and mindreading ability throughout development. Communicative interactions with young children are likely to limit the requirements for both; and as children gain greater general knowledge (i.e. access to cultural and familial common ground), communicative interactions with them may begin to exert greater demands on their inferences.

Finally, there is a role for mindreading in determining whether common ground has been successfully established, and whether it is being maintained. We do not assume a state of common ground with anyone and everyone – we need some kind of evidence. In the typical state of joint attention in infants, the individuals involved look repeatedly from the object to each other and back, layering increasing levels of assurance that the state is entirely shared. Minimal re-

sponses in spoken language assure the speaker that their audience is keeping up with them and understands what they are saying. Design elements in text messaging applications allow senders to see when their messages have been read. Again, this confirmation may become progressively more complex with development. In infant joint attention, confirmation of shared visual perspective (and therefore level-1 visual perspective-taking) is sufficient to confirm that attention is shared. With less visually apparent mental states (like knowledge), linguistic assurance may be necessary. At the extreme end of the scale with very complex content in common ground, considerable effort may be necessary to establish that common ground has been achieved: for instance, an exam or essay to confirm that a student has understood the content of a textbook.

A common ground of communicative intentions therefore does not remove the need for mindreading in communication, but *does* remove the need for high levels of recursive mindreading to underlie every ostensive utterance or behaviour. This account solves the plausibility problem by explaining informative intentions in terms of behaviour that is uncontroversially present in humans and adults, clarifies the role of mindreading in communication, and offers an account of communication that does not suffer from the problem of coordinated attack. Importantly, it does not invalidate the possibility of describing communicative intentions using metarepresentation. This point, and the significance of it, will be explained in the next section.

8.3.4 A more complete account

Minimalist accounts of Griceanism attempt to find ways in which the metarepresentational demands of communicative intentions can be reduced. This involves attempts to reduce the demands of the mental states involved (Breheny 2006) or the number of levels of metarepresentation required (Moore 2014; Moore 2016a; Tomasello 2008), resulting in impoverished accounts that fail to capture crucial components of ostensive behaviour.

By contrast, the common ground account does not require any elimination of any part of the metarepresentational account – in fact, quite the opposite, since the common ground account argues that the metarepresentational account is incomplete and fails to fully capture the openness of a state of common ground.

The iterations of the recursive regress can still be used to characterise the possible implicit knowledge of each individual in the state of jointness – that is, knowledge that logically follows from each person's mental states, but that they are not necessarily required to access or use in comprehending the situation. To illustrate this, consider that anyone who is able to count is capable of representing that 8,965 follows 8,964, but does not need to represent this fact explicitly

in order to be able to count (Wilby 2010). Similarly, if I really think about it, I may end up contemplating each stage of the regress of Meredith's and my interaction over the keys, but I need not do so in order to be satisfied that "we know that Meredith's keys are here."

As mentioned in Section 8.2.3, an explanation of ostension as otherwise inexplicable behaviour does not imply that the individuals involved cannot have metarepresentations of the informative intention, the recognition of this intention, the recognition of this recognition, and so on. When Johnny notices my ostensive behaviour, he may think, "That letter's in a weird place. She must be trying to tell me something." Later, when I get home, I may notice that the letter is gone and say, "Good, you understood my message then." In both cases, we are reflecting on the presence of my informative intention and Johnny's correct understanding of that informative intention. In this way, the metarepresentational description of communicative intentions is one way of characterising the implications of a state of common ground, but is not necessarily descriptive of the psychological states of the participants.

Wilby (2010) distinguishes between two different articulations of the problem of the infinite regress. Articulations of the plausibility problem of Griceanism fall into the same two camps. The first articulation, implicitly underlying minimalist accounts of Griceanism, identifies the problem as being a tension between the limitations of human cognition and an idealised but psychologically unrealistic description of communicative intentions. The solution to this problem is to argue that the description of communicative intentions is an abstract philosophical extrapolation that does not play a role in actual human cognition, and that simpler processes will suffice. Attempts to articulate these simpler processes, however, are not successful.

As demonstrated by the problem of coordinated attack, a more fundamental concern is a tension between the limitations of human cognition (that is, an inability to comprehend an infinite regress) and an account of common ground that successfully captures the property of jointness in this state. Having developed this account by drawing on the psychological primacy of a second-person state, the iterative account *does* become simply an extrapolation of the possible implicit knowledge of the individuals involved.

In summary, the common ground account loses none of the features of the metarepresentational account, but offers a more complete account of ostension, with a more clearly articulated role for mindreading. This account raises a substantial set of new research questions, both empirical and theoretical. In the final section, I will make some suggestions for new avenues for research on ostension.

8.4 Suggestions for future research

The definition of ostensive behaviour as actions that do not have another more plausible explanation invites empirical research investigating whether this is in fact what causes people to infer communicative intent. Much like pareidolia (the tendency to see faces in objects) and apophenia (the mistaken identification of patterns in randomness), it seems plausible that the ready assumption of communicative intent is a central feature of human cognition. (In fact, humans are so given to reading attempts to communicate in otherwise inexplicable occurrences that we will readily attribute supernatural "signs" to coincidences like oddly-timed lightning strikes.)

It is an empirical question why this tendency exists: it could be the result of a lifetime of training, or of an innate, species-characteristic behaviour (Csibra and Gergely 2011). Probing the contexts in which the inference from unexpected behaviour to communication occurs and does not occur would be a useful way to establish the criteria for a behaviour to be interpreted as ostensive. Research on the efficiency, effort and conscious awareness of attributions of communicative intent would similarly help to establish whether this is a conscious or effortful inference, or whether it is automatic, spontaneous, or otherwise rapid and involuntary. Ongoing research on the recognition of communicative behaviour in infants is highly relevant to an understanding of ostension.

As discussed in Section 8.3.3, mindreading plays a role in the decision of what information to move into common ground, what cues to include in an informative intention to successfully achieve this, and inferring the meaning of an informative intention. Throughout development, and with greater degrees of shared context between interlocutors, the extent of the inference required based on the evidence may increase. Corpus studies of child-directed speech may be informative of the development and pacing of more sophisticated communicative interactions throughout development. There is also room to draw on psycholinguistic research on interactive alignment to understand how priming, memory and other attentional processes may reduce the need for open-ended inferences about mental states in everyday communication, further clarifying the specific role of mindreading (Apperly 2018; Garrod and Pickering 2004; Garrod and Pickering 2007; Pickering and Garrod 2004).

Another developmental question is whether children's communication follows the trajectory of more concrete to more abstract communication outlined in Section 8.2.2 – that is, with early communication focusing on shared attention to objects and events in the immediate environment, and later incorporating abstract concepts, objects and events distant in time and space,

and eventually asynchronous communication.

The utility of focusing on second-person experience in social cognition is receiving increasing attention (Eilan 2005; Eilan 2016; Siposova and Carpenter 2019; Schilbach et al. 2013). This direction offers new opportunities for investigating the broad range of behaviours within common ground and shared attention, including which levels of social attention are achievable by humans of various ages, and by non-human animals. This provides a new theoretical avenue for ostension, in addition to the suggestions for research mentioned above: there is considerable theoretical work to be done in mapping the conceptual territory of different "common ground" accounts of communication (Sperber and Wilson 1986; Eilan 2005; Tomasello 2008; Peacocke 2005; Clark and Brennan 1991); the points of agreement and contention between them; and the overlap with the growing literature on second-person social cognition.

8.5 Chapter summary

In this chapter, I have argued that a reframing of ostensive-inferential communication more successfully explains how ostension is achieved, and offers substantial advantages and opportunities for future research. I explained the criticism of ostension on the grounds of plausibility, describing minimalist accounts of Griceanism developed in response to this criticism, and critiquing the shortfalls of these accounts.

I argued that shared attention and common ground provide a better reformulation of ostension, losing none of the essential features of Gricean communicative intention, but characterising communication in terms of abilities that are uncontroversially present in infants and carry greater psychological plausibility in everyday interactions in adults.

I discussed the problem of infinite regress, and how conceptualising "mutualness" as a psychological primitive avoids an infinite regress, providing a more complete account of common ground. I offered a characterisation of a primitive "we-state" as avoiding the necessity to represent the mental states of another individual as separate from one's own, which distinguishes betweeen this behaviour and mindreading, and offers a more constrained role for mindreading in communication. I explained how the common ground account differs from minimalist accounts of Griceanism, instead offering an explanation for ostension that is both cognitively plausible and more complete than a metarepresentational account. Finally, I offered suggestions for redirected research on ostension based on this account.

Chapter 9

Summary and conclusions

This thesis has explored the role of mindreading in language evolution by investigating rapid and involuntary perspective-taking in the Dot Perspective Task. After outlining the structure of the thesis in Chapter 1, Chapter 2 reviewed the current literature on evolutionary linguistics, which has achieved remarkable insights on the evolution of linguistic structure as the product of cultural transmission. This research, though, has not focused on how signal-meaning mappings emerge, leaving an important gap in this understanding. Modelling and empirical work suggests that there are limited routes by which signal-meaning mappings can evolve independently and later become connected, but only one route by which they may emerge simultaneously, as seen in human communication: the signalling of signalhood, or displaying of communicative intentions. This *ostensive-inferential* communication, which relies on the display and interpretation of communicative intentions, is heavily reliant on mindreading. Mindreading, then, is crucial to understanding the evolution of language.

The plausibility of the ostensive-inferential account has been questioned, on the grounds that it requires a mindreading capacity which has not uncontroversially been found in children, and that appears advanced and difficult even for adults. Chapter 3 reviewed the literature on mindreading: in children, across cultures, across species, and adults, establishing that the current empirical picture is mixed and complex. Various theoretical accounts attempt to make sense of this literature in different ways, positing one or two different mindreading systems; an innate or learned mindreading capacity; and the presence of either true mindreading or "submentalising" abilites in behaviour that appear to require mindreading. The ostensive-inferential account relies on a particular account of mindreading: one unified system that does not entail submentalising. Studies of rapid and involuntary perspective-taking in adults offer a useful opportunity to test the predictions of various accounts of mindreading, and to test the limits of rapid perspective-taking

in adults, both of which are informative to the ostensive-inferential model of communication.

Chapters 4 to 7 presented a series of experiments designed to test these predictions and limits using the Dot Perspective Task (DPT), a paradigm that appears to demonstrate investigate rapid and involuntary perspective-taking. Chapter 4 compared rapid and involuntary perspective-taking of humanoid avatars to visually similar arrows, finding a null result. The apparent explanation for this null result – the lack of task requirements prompting participants to consider the arrows or avatars as relevant to the task – is corroborated by similar findings in other "uncued" tasks in the DPT literature. A review of this literature established a methodological inconsistency in the DPT literature that appears to underlie contradictory results, and established how this inconsistency could be operationalised to produce informative results in an adapted DPT task comparing submentalising and mentalising explanations of the effect.

Chapter 5 presented this adapted DPT task, as well as four follow-up experiments. In Experiment 1, DPT variants with three different sets of task demands were compared: an "explicit" task requiring participants to take the avatar's perspective throughout the task; an "implicit" task requiring no explicit perspective-taking, but with cues that may have prompted the participants to consider the avatar's perspective as salient; and an "uncued" task that made no mention of the avatars at all. While the explicit task found an altercentric effect best explained by mentalising, not submentalising, and the uncued task found the expected null result, the implicit task returned an unexpected null result. Experiments 2 to 5 investigated possible reasons for this null, establishing that the appearance of the stimuli (Lego characters rather than the humanoid avatars used in much of the DPT literature) appears have an effect on results. These results, collectively, suggest that the effect in the DPT is best explained by perspective-taking rather than directional orienting, and that it is not automatic (that is, reflexively determined by the presence of an agent in the scene), but is rather spontaneous (that is, prompted by top-down information such as task demands and characteristics). In particular, the requirement to take the avatar's perspective throughout the task appears crucial.

Chapter 6 extended the explicit task in Chapter 5 to test the predictions of two different accounts of the cause of the altercentric effect: processing costs (that is, a delay caused by processing conflicting perspectives) and preferential attention (that is, a delay caused by preferential attention to the more "salient" balls positioned in a joint visual field). Two experiments returned null results, suggesting a limitation to the complexity of spontaneous perspective-taking in the DPT. In light of these and other null results presented in this thesis, Chapter 7 discussed the statistical and institutional mechanisms underlying a low replication rate in fields that rely on

statistical inference, and how these problems pertain to the DPT literature. I suggested a range of best practices that should be implemented in future DPT research in order to ensure more empirically robust results, including increased sample sizes, pre-registered experiments, and open materials and analyses.

Finally, Chapter 8 discussed the implications of this empirical work for the ostensive-inferential model of communication, arguing that evidence against submentalising and two-systems interpretations of the DPT, and evidence of rapid and highly successful perspective-taking both provide tentative, albeit indirect, evidence in favour of the ostensive-inferential model. I suggested that future empirical research on ostension should focus on joint attention, rather than perspective-taking, offering an alternative analysis of communicative intentions that relies on joint attention rather than recursive mindreading. I argued that this account offers a more complete and cognitively plausible account of communicative intentions than either the metarepresentational account or minimalist alternatives. Finally, I suggested possible avenues for future research.

References

Akhtar, N. and M. Tomasello (1996). "Two-Year-Olds Learn Words for Absent Objects and Actions". In: *British Journal of Developmental Psychology* 14.1, pp. 79–93. DOI: 10.1111/j.2044-835X.1996.tb00695.x.

Alsheikh-Ali, A. A., W. Qureshi, M. H. Al-Mallah, and J. P. A. Ioannidis (2011). "Public Availability of Published Research Data in High-Impact Journals". In: *PLoS ONE* 6.9. Ed. by I. Boutron, e24357. DOI: 10.1371/journal.pone.0024357.

Amrhein, V. and S. Greenland (2018). "Remove, Rather than Redefine, Statistical Significance". In: *Nature Human Behaviour* 2.1, pp. 4–4. DOI: 10.1038/s41562-017-0224-0.

Anderson, C. J. et al. (2016). "Response to Comment on "Estimating the Reproducibility of Psychological Science"". In: *Science* 351.6277, pp. 1037–1037. DOI: 10.1126/science.aad9163.

Anderson, D. K., C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, K. Welch, and A. Pickles (2007). "Patterns of Growth in Verbal Abilities among Children with Autism Spectrum Disorder." In: *Journal of Consulting and Clinical Psychology* 75.4, pp. 594–604. DOI: 10.1037/0022-006X.75.4.594.

Apicella, C. L. and H. C. Barrett (2016). "Cross-Cultural Evolutionary Psychology". In: *Current Opinion in Psychology* 7, pp. 92–97. DOI: 10.1016/j.copsyc.2015.08.015.

Apperly, I. A. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. OCLC: ocn432998292. Hove [East Sussex] ; New York: Psychology Press.

— (2018). "Mindreading and Psycholinguistic Approaches to Perspective Taking: Establishing Common Ground". In: *Topics in Cognitive Science* 10.1, pp. 133–139. DOI: 10.1111/tops.12308.

Apperly, I. A. and S. A. Butterfill (2009). "Do Humans Have Two Systems to Track Beliefs and Belief-like States?" In: *Psychological Review* 116.4, pp. 953–970. DOI: 10.1037/a0016923.

Apperly, I. A., D. Samson, and G. W. Humphreys (2009). "Studies of Adults Can Inform Accounts of Theory of Mind Development." In: *Developmental Psychology* 45.1, pp. 190–201. DOI: 10.1037/a0014098.

Astington, J. W. and J. A. Baird, eds. (2005). *Why Language Matters for Theory of Mind*. Oxford University Press.

Astuti, R. (2015). "Implicit and Explicit Theory of Mind". In: *Anthropology of This Century* 13, pp. 636–650.

Avis, J. and P. L. Harris (1991). "Belief-Desire Reasoning among Baka Children: Evidence for a Universal Conception of Mind". In: *Child Development*, p. 9.

Baayen, H. and P. Milin (2010). "Analyzing Reaction Times". In: *International Journal of Psychological Research* 3.2, p. 12. DOI: 10.21500/20112084.807.

Baillargeon, R. (1994). "How Do Infants Learn about the Physical World?" In: *Current Directions in Psychological Science*.

Baker, L. J., D. T. Levin, and M. M. Saylor (2016). "The Extent of Default Visual Perspective Taking in Complex Layouts." In: *Journal of Experimental Psychology: Human Perception and Performance* 42.4, pp. 508–516. DOI: 10.1037/xhp0000164.

Baker, M. (2016). "A Nature Survey Lifts the Lid on How Researchers View the 'Crisis' Rocking Science and What They Think Will Help." In: p. 3.

Bakker, M., A. van Dijk, and J. M. Wicherts (2012). "The Rules of the Game Called Psychological Science". In: *Perspectives on Psychological Science* 7.6, pp. 543–554. DOI: 10.1177/1745691612459060.

Bakker, M. and J. M. Wicherts (2011). "The (Mis)Reporting of Statistical Results in Psychology Journals". In: *Behavior Research Methods* 43.3, pp. 666–678. DOI: 10.3758/s13428-011-0089-5.

Baldwin, D. A. and L. J. Moses (2001). "Links between Social Understanding and Early Word Learning: Challenges to Current Accounts". In: *Social Development* 10.3, pp. 309–329. DOI: 10.1111/1467-9507.00168.

Bargh, J. A. (1994). "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition". In: *Handbook of Social Cognition: Basic Processes; Applications*. Ed. by R. S. Wyer and T. K. Srull. Lawrence Erlbaum.

Baron-Cohen, S. (1988). "Social and Pragmatic Deficits in Autism: Cognitive or Affective?" In: *Journal of Autism and Developmental Disorders* 18.3, pp. 379–402. DOI: 10.1007/BF02212194.

— (1995). "Mindblindness: An Essay on Autism and Theory of Mind". In: *Learning Development and Conceptual Change*.

Baron-Cohen, S., D. A. Baldwin, and M. Crowson (1997). "Do Children with Autism Use the Speaker's Direction of Gaze Strategy to Crack the Code of Language?" In: *Child Development* 68.1, pp. 48–57.

Baron-Cohen, S., A. M. Leslie, and U. Frith (1985). "Does the Autistic Child Have a "Theory of Mind" ?" In: *Cognition* 21.1, pp. 37–46. DOI: 10.1016/0010-0277(85)90022-8.

Barrett, H. C. et al. (2013). "Early False-Belief Understanding in Traditional Non-Western Societies". In: *Proceedings of the Royal Society B: Biological Sciences* 280.1755, pp. 20122654–20122654. DOI: 10.1098/rspb.2012.2654.

Bartlett, F. C. (1932). *Remembering*. Cambridge University Press.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). "Fitting Linear Mixed-Effects Models Using Lme4". In: *Journal of Statistical Software* 67.1. DOI: 10.18637/jss.v067.i01.

Battich, L. and B. Geurts (2018). "Joint Attention Is Not Just Perception". In: *Preprint*, p. 14.

Bayliss, A. P. and S. P. Tipper (2006). "Predictive Gaze Cues and Personality Judgments: Should Eye Trust You?" In: *Psychological Science* 17.6, pp. 514–520. DOI: 10.1111/j.1467-9280.2006.01737.x.

Begley, C. G. and L. M. Ellis (2012). "Raise Standards for Preclinical Cancer Research: Drug Development". In: *Nature* 483.7391, pp. 531–533. DOI: 10.1038/483531a.

Benjamin, D. J. et al. (2018). "Redefine Statistical Significance". In: *Nature Human Behaviour* 2.1, pp. 6–10. DOI: 10.1038/s41562-017-0189-z.

Berko, J. (1958). "The Child's Learning of English Morphology". In: *Word* 14.2-3, pp. 150–177.

Berwick, R. C., K. Okanoya, G. J. Beckers, and J. J. Bolhuis (2011). "Songs to Syntax: The Linguistics of Birdsong". In: *Trends in Cognitive Sciences* 15.3, pp. 113–121. DOI: 10.1016/j.tics.2011.01.002.

Bloom, P. and T. P. German (2000). "Two Reasons to Abandon the False Belief Task as a Test of Theory of Mind". In: *Cognition* 77.1, B25–B31. DOI: 10.1016/S0010-0277(00)00096-2.

Blythe, R. A. and T. C. Scott-Phillips (2014). "Simulating the Real Origins of Communication". In: *PLoS ONE* 9.11. Ed. by S. A. White, e113636. DOI: 10.1371/journal.pone.0113636.

Bradbury, J. W. and S. L. Vehrencamp (2011). *Principles of Animal Communication*. Sunderland, MA: Sinauer Associates.

Brandt, M. J. et al. (2014). "The Replication Recipe: What Makes for a Convincing Replication?" In: *Journal of Experimental Social Psychology* 50, pp. 217–224. DOI: 10.1016/j.jesp.2013.10.005.

Breheny, R. (2006). "Communication and Folk Psychology". In: *Mind and Language* 21.1, pp. 74–107. DOI: 10.1111/j.1468-0017.2006.00307.x.

Brighton, H., K. Smith, and S. Kirby (2005). "Language as an Evolutionary System". In: *Physics of Life Reviews* 2.3, pp. 177–226. DOI: 10.1016/j.plrev.2005.06.001.

Brooks, R. and A. N. Meltzoff (2015). "Connecting the Dots from Infancy to Childhood: A Longitudinal Study Connecting Gaze Following, Language, and Explicit Theory of Mind". In: *Journal of Experimental Child Psychology* 130, pp. 67–78. DOI: 10.1016/j.jecp.2014.09.010.

Brown-Schmidt, S. (2009). "The Role of Executive Function in Perspective Taking during Online Language Comprehension". In: *Psychonomic Bulletin & Review* 16.5, pp. 893–900. DOI: 10.3758/PBR.16.5.893.

Bukowski, H., J. K. Hietanen, and D. Samson (2015). "From Gaze Cueing to Perspective Taking: Revisiting the Claim That We Automatically Compute Where or What Other People Are Looking At". In: *Visual Cognition* 23.8, pp. 1020–1042. DOI: 10.1080/13506285.2015.1132804.

Buttelmann, D., F. Buttelmann, M. Carpenter, J. Call, and M. Tomasello (2017). "Great Apes Distinguish True from False Beliefs in an Interactive Helping Task". In: *PLOS ONE* 12.4. Ed. by J. Kaminski, e0173793. DOI: 10.1371/journal.pone.0173793.

Buttelmann, D., M. Carpenter, and M. Tomasello (2009). "Eighteen-Month-Old Infants Show False Belief Understanding in an Active Helping Paradigm". In: *Cognition* 112.2, pp. 337–342. DOI: 10.1016/j.cognition.2009.05.006.

Buttelmann, D., H. Over, M. Carpenter, and M. Tomasello (2014). "Eighteen-Month-Olds Understand False Beliefs in an Unexpected-Contents Task". In: *Journal of Experimental Child Psychology* 119, pp. 120–126. DOI: 10.1016/j.jecp.2013.10.002.

Butterworth, G. (1995). "Origins of Minds in Perception and Action". In: *Joint Attention: Its Origins and Role in Development*. Ed. by C. Moore and P. J. Dunham. Erlbaum.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò (2013). "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience". In: *Nature Reviews Neuroscience* 14.5, pp. 365–376. DOI: 10.1038/nrn3475.

Caldwell, C. A. and K. Smith (2012). "Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties". In: *PLoS ONE* 7.8. Ed. by A. Mesoudi, e43807. DOI: 10.1371/journal.pone.0043807.

Callaghan, T., P. Rochat, A. Lillard, M. L. Claux, H. Odden, S. Itakura, S. Tapanya, and S. Singh (2005). "Synchrony in the Onset of Mental-State Reasoning: Evidence From Five Cultures". In: *Psychological Science* 16.5, pp. 378–384. DOI: 10.1111/j.0956-7976.2005.01544.x.

Camerer, C. F. et al. (2018). "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015". In: *Nature Human Behaviour* 2.9, pp. 637–644. DOI: 10.1038/s41562-018-0399-z.

Camerer, C. F. et al. (2016). "Evaluating Replicability of Laboratory Experiments in Economics". In: p. 5.

Campbell, J. (2005). "Joint Attention and Common Knowledge". In: *Joint Attention: Communication and Other Minds*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199245635.001.0001.

Capozzi, F., A. Cavallo, T. Furlanetto, and C. Becchio (2014). "Altercentric Intrusions from Multiple Perspectives: Beyond Dyads". In: *PLoS ONE* 9.12. Ed. by C. Urgesi, e114210. DOI: 10.1371/journal.pone.0114210.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford University Press.

— (2013). "Mindreading in Infancy". In: *Mind & Language* 28.2, pp. 141–172. DOI: 10.1111/mila.12014.

— (2017). "Mindreading in Adults: Evaluating Two-Systems Views". In: *Synthese* 194.3, pp. 673–688. DOI: 10.1007/s11229-015-0792-3.

Carston, R. (2002). "Linguistic Meaning, Communicated Meaning and Cognitive Pragmatics". In: *Mind and Language* 17.1&2, pp. 127–148. DOI: 10.1111/1468-0017.00192.

— (2013). "Word Meaning, What Is Said, and Explicature". In: *What Is Said and What Is Not*. Stanford: Csli Publications, p. 28.

Catmur, C., I. Santiesteban, J. R. Conway, C. Heyes, and G. Bird (2016). "Avatars and Arrows in the Brain". In: *NeuroImage* 132, pp. 8–10. DOI: 10.1016/j.neuroimage.2016.02.021.

Chambers, C. D. (2013). "Registered Reports: A New Publishing Initiative at Cortex". In: *Cortex* 49.3, pp. 609–610. DOI: 10.1016/j.cortex.2012.12.016.

Chater, N. and M. H. Christiansen (2010). "Language Acquisition Meets Language Evolution". In: *Cognitive Science* 34.7, pp. 1131–1157. DOI: 10.1111/j.1551-6709.2009.01049.x.

Christiansen, M. H. and N. Chater (2008). "Language as Shaped by the Brain". In: *Behavioral and Brain Sciences* 31.05. DOI: 10.1017/S0140525X08004998.

Claidière, N., K. Smith, S. Kirby, and J. Fagot (2014). "Cultural Evolution of Systematically Structured Behaviour in a Non-Human Primate". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1797, pp. 20141541–20141541. DOI: 10.1098/rspb.2014.1541.

Clark, H. H. and S. E. Brennan (1991). "Grounding in Communication." In: *Perspectives on Socially Shared Cognition.* Ed. by L. B. Resnick, J. M. Levine, and S. D. Teasley. Washington: American Psychological Association, pp. 127–149. DOI: 10.1037/10096-006.

Cohen, J. (1990). "Things I Have Learned (So Far)". In: *American Psychologist*, p. 9.

Cole, G., M. Atkinson, A. D'Souza, and D. Smith (2017). "Spontaneous Perspective Taking in Humans?" In: *Vision* 1.2, p. 17. DOI: 10.3390/vision1020017.

Cole, G., M. Atkinson, A. T. Le, and D. T. Smith (2016). "Do Humans Spontaneously Take the Perspective of Others?" In: *Acta Psychologica* 164, pp. 165–168. DOI: 10.1016/j.actpsy.2016.01.007.

Collins, S. (2004). "Vocal Fighting and Flirting: The Functions of Birdsong". In: *Nature's Music: The Science of Birdsong.* Academic Press.

Coltheart, M. (1999). "Modularity and Cognition". In: *Trends in Cognitive Sciences* 3.3, pp. 115–120. DOI: 10.1016/S1364-6613(99)01289-9.

Conway, J. R., D. Lee, M. Ojaghi, C. Catmur, and G. Bird (2017). "Submentalizing or Mentalizing in a Level 1 Perspective-Taking Task: A Cloak and Goggles Test." In: *Journal of Experimental Psychology: Human Perception and Performance* 43.3, pp. 454–465. DOI: 10.1037/xhp0000319.

Cooper, H., K. DeNeve, and K. Charlton (1997). "Finding the Missing Science: The Fate of Studies Submitted for Review by a Human Subjects Committee". In: *Psychological Methods* 2, pp. 447–452.

Coursol, A. and E. E. Wagner (1986). "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias". In: *Professional Psychology: Research and Practice* 17, pp. 136–137.

Csibra, G. (2010). "Recognizing Communicative Intentions in Infancy". In: *Mind & Language* 25.2, pp. 141–168. DOI: 10.1111/j.1468-0017.2009.01384.x.

Csibra, G. and G. Gergely (2009). "Natural Pedagogy". In: *Trends in Cognitive Sciences* 13.4, pp. 148–153. DOI: 10.1016/j.tics.2009.01.005.

— (2011). "Natural Pedagogy as Evolutionary Adaptation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1567, pp. 1149–1157.

Culbertson, J. and D. Adger (2014). "Language Learners Privilege Structured Meaning over Surface Frequency". In: *Proceedings of the National Academy of Sciences* 111.16, pp. 5842–5847. DOI: 10.1073/pnas.1320525111.

Culbertson, J. and E. L. Newport (2015). "Harmonic Biases in Child Learners: In Support of Language Universals". In: *Cognition* 139, pp. 71–82. DOI: 10.1016/j.cognition.2015.02.007.

Culbertson, J., P. Smolensky, and G. Legendre (2012). "Learning Biases Predict a Word Order Universal". In: *Cognition* 122.3, pp. 306–329. DOI: 10.1016/j.cognition.2011.10.017.

de Boer, B. (2000). "Self-Organization in Vowel Systems". In: *Journal of Phonetics* 28.4, pp. 441–465. DOI: 10.1006/jpho.2000.0125.

— (2001). *The Origins of Vowel Systems*. Oxford University Press.

De Giacomo, A. and E. Fombonne (1998). "Parental Recognition of Developmental Abnormalities in Autism". In: *European Child and Adolescent Psychiatry* 7.3, pp. 131–136.

de Marchena, A., I.-M. Eigsti, A. Worek, K. E. Ono, and J. Snedeker (2011). "Mutual Exclusivity in Autism Spectrum Disorders: Testing the Pragmatic Hypothesis". In: *Cognition* 119.1, pp. 96–113. DOI: 10.1016/j.cognition.2010.12.011.

de Villiers, J. G. and J. E. Pyers (2002). "Complements to Cognition: A Longitudinal Study of the Relationship between Complex Syntax and False-Belief-Understanding". In: *Cognitive Development* 17.1, pp. 1037–1060. DOI: 10.1016/S0885-2014(02)00073-4.

Dingemanse, M., F. Torreira, and N. J. Enfield (2013). "Is "Huh?" A Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items". In: *PLoS ONE* 8.11. Ed. by J. J. Bolhuis, e78273. DOI: 10.1371/journal.pone.0078273.

Drayton, L. A., L. R. Santos, and A. Baskin-Sommers (2018). "Psychopaths Fail to Automatically Take the Perspective of Others". In: *Proceedings of the National Academy of Sciences* 115.13, pp. 3302–3307. DOI: 10.1073/pnas.1721903115.

Duh, S., J. H. Paik, P. H. Miller, S. C. Gluck, H. Li, and I. Himelfarb (2016). "Theory of Mind and Executive Function in Chinese Preschool Children." In: *Developmental Psychology* 52.4, pp. 582–591. DOI: 10.1037/a0040068.

Dunn, M., S. J. Greenhill, S. C. Levinson, and R. D. Gray (2011). "Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals". In: *Nature* 473.7345, pp. 79–82. DOI: 10.1038/nature09923.

Eigsti, I.-M., A. B. de Marchena, J. M. Schuh, and E. Kelley (2011). "Language Acquisition in Autism Spectrum Disorders: A Developmental Review". In: *Research in Autism Spectrum Disorders* 5.2, pp. 681–691. DOI: 10.1016/j.rasd.2010.09.001.

Eilan, N. (2005). "Joint Attention, Communication, and Mind". In: *Joint Attention: Communication and Other Minds*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199245635.001.0001.

— (2016). "Joint Attention and the Second Person". In: *Preprint*.

Elekes, F., M. Varga, and I. Király (2016). "Evidence for Spontaneous Level-2 Perspective Taking in Adults". In: *Consciousness and Cognition* 41, pp. 93–103. DOI: 10.1016/j.concog.2016.02.010.

Etz, A. and J. Vandekerckhove (2016). "A Bayesian Perspective on the Reproducibility Project: Psychology". In: *PLOS ONE* 11.2. Ed. by D. Marinazzo, e0149794. DOI: 10.1371/journal.pone.0149794.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi (1995). *Reasoning about Knowledge*. MIT Press.

Fanelli, D. (2010). ""Positive" Results Increase Down the Hierarchy of the Sciences". In: *PLoS ONE* 5.4, p. 10.

Farroni, T., M. H. Johnson, M. Brockbank, and F. Simion (2000). "Infants' Use of Gaze Direction to Cue Attention: The Importance of Perceived Motion". In: *Visual Cognition* 7.6, pp. 705–718. DOI: 10.1080/13506280050144399.

Fay, N., S. Garrod, L. Roberts, and N. Swoboda (2010). "The Interactive Evolution of Human Communication Systems". In: *Cognitive Science* 34.3, pp. 351–386. DOI: 10.1111/j.1551-6709.2009.01090.x.

Fedzechkina, M., T. F. Jaeger, and E. L. Newport (2012). "Language Learners Restructure Their Input to Facilitate Efficient Communication". In: *Proceedings of the National Academy of Sciences* 109.44, pp. 17897–17902. DOI: 10.1073/pnas.1215776109.

Fehér, O., H. Wang, S. Saar, P. P. Mitra, and O. Tchernichovski (2009). "De Novo Establishment of Wild-Type Song Culture in the Zebra Finch". In: *Nature* 459.7246, pp. 564–568. DOI: 10.1038/nature07994.

Ferguson, C. (2009). "An Effect Size Primer: A Guide for Clinicians and Researchers." In: *Professional Psychology: Research and Practice* 40.5, pp. 532–538. DOI: 10.1037/a0015808.

Ferguson, C. and M. T. Brannick (2012). "Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses." In: *Psychological Methods* 17.1, pp. 120–128. DOI: 10.1037/a0024445.

Ferguson, H., I. A. Apperly, and J. E. Cane (2017). "Eye Tracking Reveals the Cost of Switching between Self and Other Perspectives in a Visual Perspective-Taking Task". In: *Quarterly Journal of Experimental Psychology* 70.8, pp. 1646–1660. DOI: 10.1080/17470218.2016.1199716.

Ferguson, H., V. E. A. Brunsdon, and E. E. F. Bradford (2018). "Age of Avatar Modulates the Altercentric Bias in a Visual Perspective-Taking Task: ERP and Behavioral Evidence". In: *Cognitive, Affective, & Behavioral Neuroscience* 18.6, pp. 1298–1319. DOI: 10.3758/s13415-018-0641-1.

Flavell, J. H. (1977). "The Development of Knowledge about Visual Perception". In: *Nebraska Symposium on Motivation*.

Flavell, J. H., E. R. Flavell, and F. L. Green (1983). "Development of the Appearance-Reality Distinction". In: *Cognitive Psychology* 15.1, pp. 95–120. DOI: 10.1016/0010-0285(83)90005-1.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press.

Francis, G. (2012a). "The Psychology of Replication and Replication in Psychology". In: *Perspectives on Psychological Science* 7.6, pp. 585–594. DOI: 10.1177/1745691612459520.

— (2012b). "Too Good to Be True: Publication Bias in Two Prominent Studies from Experimental Psychology". In: *Psychonomic Bulletin & Review* 19.2, pp. 151–156. DOI: 10.3758/s13423-012-0227-9.

Franco, A., N. Malhotra, and G. Simonovits (2016). "Underreporting in Psychology Experiments: Evidence From a Study Registry". In: *Social Psychological and Personality Science* 7.1, pp. 8–12. DOI: 10.1177/1948550615598377.

Frank, M. C. et al. (2017). "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building". In: *Infancy* 22.4, pp. 421–435. DOI: 10.1111/infa.12182.

Freedman, L. P., I. M. Cockburn, and T. S. Simcoe (2015). "The Economics of Reproducibility in Preclinical Research". In: *PLOS Biology* 13.6, e1002165. DOI: 10.1371/journal.pbio.1002165.

Freundlieb, M., Á. M. Kovács, and N. Sebanz (2016). "When Do Humans Spontaneously Adopt Another's Visuospatial Perspective?" In: *Journal of Experimental Psychology: Human Perception and Performance* 42.3, pp. 401–412. DOI: 10.1037/xhp0000153.

Freundlieb, M., Á. M. Kovács, and N. Sebanz (2018). "Reading Your Mind While You Are Reading—Evidence for Spontaneous Visuospatial Perspective Taking During a Semantic Categorization Task". In: *Psychological Science* 29.4, pp. 614–622. DOI: 10.1177/0956797617740973.

Freundlieb, M., N. Sebanz, and Á. M. Kovács (2017). "Out of Your Sight, out of My Mind: Knowledge about Another Person's Visual Access Modulates Spontaneous Visuospatial Perspective-Taking." In: *Journal of Experimental Psychology: Human Perception and Performance* 43.6, pp. 1065–1072. DOI: 10.1037/xhp0000379.

Frischen, A., A. P. Bayliss, and S. P. Tipper (2007). "Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences." In: *Psychological Bulletin* 133.4, pp. 694–724. DOI: 10.1037/0033-2909.133.4.694.

Fugelsang, J. A., C. B. Stein, A. E. Green, and K. N. Dunbar (2004). "Theory and Data Interactions of the Scientific Mind: Evidence From the Molecu..." In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 58.2, p. 10.

Furlanetto, T., C. Becchio, D. Samson, and I. A. Apperly (2016). "Altercentric Interference in Level 1 Visual Perspective Taking Reflects the Ascription of Mental States, Not Submentalizing." In: *Journal of Experimental Psychology: Human Perception and Performance* 42.2, pp. 158–163. DOI: 10.1037/xhp0000138.

Furlanetto, T., A. Cavallo, V. Manera, B. Tversky, and C. Becchio (2013). "Through Your Eyes: Incongruence of Gaze and Action Increases Spontaneous Perspective Taking". In: *Frontiers in Human Neuroscience* 7. DOI: 10.3389/fnhum.2013.00455.

Gagne, D. L. and M. Coppola (2017). "Visible Social Interactions Do Not Support the Development of False Belief Understanding in the Absence of Linguistic Input: Evidence from Deaf Adult Homesigners". In: *Frontiers in Psychology* 8. DOI: 10.3389/fpsyg.2017.00837.

Galantucci, B., C. Kroos, and T. Rhodes (2010). "The Effects of Rapidity of Fading on Communication Systems". In: *Interaction Studies* 11.1.

Gardner, M. R., A. Bileviciute, and C. Edmonds (2018a). "Implicit Mentalising during Level-1 Visual Perspective-Taking Indicated by Dissociation with Attention Orienting". In: *Vision* 2.1, p. 3. DOI: 10.3390/vision2010003.

Gardner, M. R., Z. Hull, D. Taylor, and C. J. Edmonds (2018b). "'Spontaneous' Visual Perspective-Taking Mediated by Attention Orienting That Is Voluntary and Not Reflexive". In: *Quarterly Journal of Experimental Psychology* 71.4, pp. 1020–1029. DOI: 10.1080/17470218.2017.1307868.

Garrod, Simon, Fay, Nicolas, Rogers, Shane, Walker, Bradley, and Swoboda, Nik (2010). "Can Iterated Learning Explain the Emergence of Graphical Symbols?" In: *Interaction Studies* 11.1, pp. 33–50. DOI: 10.1075/is.11.1.04gar.

Garrod, S., N. Fay, J. Lee, J. Oberlander, and T. MacLeod (2007). "Foundations of Representation: Where Might Graphical Symbol Systems Come From?" In: *Cognitive Science* 31.6, pp. 961–987. DOI: 10.1080/03640210701703659.

Garrod, S. and M. J. Pickering (2004). "Why Is Conversation so Easy?" In: *Trends in Cognitive Sciences* 8.1, pp. 8–11. DOI: 10.1016/j.tics.2003.10.016.

— (2007). *Alignment in Dialogue*. Oxford University Press. DOI: 10.1093/oxfordhb/9780198568971.013.0026.

Gelman, A. and E. Loken (2013). "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Time". In: *Department of Statistics, Columbia University*, p. 17.

Geurts, B. (2018). "Convention and Common Ground". In: *Mind & Language* 33.2, pp. 115–129. DOI: 10.1111/mila.12171.

— (2019). "What's Wrong with Gricean Pragmatics". In: *Proceedings of ExLing 2019*, p. 8.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson (2016). "Comment on "Estimating the Reproducibility of Psychological Science"". In: *Science* 351.6277, pp. 1037–1037. DOI: 10.1126/science.aad7243.

Givón, T. (1985). "Function, Structure and Language Acquisition". In: *The Crosslinguistic Study of Language Acquisition*. Vol. 2. Hillsdale, NJ: Lawrence Erlbaum, pp. 1005–1028.

Goldin-Meadow, S., W. C. So, A. Ozyurek, and C. Mylander (2008). "The Natural Order of Events: How Speakers of Different Languages Represent Events Nonverbally". In: *Proceedings of the National Academy of Sciences* 105.27, pp. 9163–9168. DOI: 10.1073/pnas.0710060105.

Gómez, J.-C. (1996). "Non-Human Primate Theories of (Non-Human Primate) Minds: Some Issues Concerning the Origins of Mind-Reading". In: *Theories of Theories of Mind*. Ed. by P. Carruthers and P. K. Smith. Cambridge: Cambridge University Press, pp. 330–343. DOI: 10.1017/CBO9780511597985.020.

Gopnik, A. (1996). "The Scientist as Child. Philosophy of Science". In: *Philosophy of Science* 63, pp. 485–514.

Gopnik, A. and H. M. Wellman (2012). "Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory." In: *Psychological Bulletin* 138.6, pp. 1085–1108. DOI: 10.1037/a0028044.

Greenberg, J. H. (1963). "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements". In: *Universals of Language*. MIT Press, pp. 73–113.

Grice, P. (1957). "Meaning". In: *The Philosophical Review* 66.3, pp. 377–388.

Griffiths, T. L., B. Christian, and M. L. Kalish (2008). "Using Category Structures to Test Iterated Learning as a Method for Identifying Inductive Biases". In: *Cognitive Science: A Multidisciplinary Journal* 32.1, pp. 68–107. DOI: 10.1080/03640210701801974.

Griffiths, T. L. and M. L. Kalish (2007). "Language Evolution by Iterated Learning With Bayesian Agents". In: *Cognitive Science* 31.3, pp. 441–480. DOI: 10.1080/15326900701326576.

Grosse, G., T. Behne, M. Carpenter, and M. Tomasello (2010). "Infants Communicate in Order to Be Understood." In: *Developmental Psychology* 46.6, pp. 1710–1722. DOI: 10.1037/a0020727.

Guo, Y., H. L. Logan, D. H. Glueck, and K. E. Muller (2013). "Selecting a Sample Size for Studies with Repeated Measures". In: p. 8.

Hanna, J. E. and M. K. Tanenhaus (2004). "Pragmatic Effects on Reference Resolution in a Collaborative Task: Evidence from Eye Movements". In: *Cognitive Science* 28.1, pp. 105–115. DOI: 10.1207/s15516709cog2801_5.

Hanna, J. E., M. K. Tanenhaus, and J. C. Trueswell (2003). "The Effects of Common Ground and Perspective on Domains of Referential Interpretation". In: *Journal of Memory and Language* 49.1, pp. 43–61. DOI: 10.1016/S0749-596X(03)00022-6.

Harris, P. L. (1996). "Desires, Beliefs, and Language". In: *Theories of Theories of Mind*. Ed. by P. Carruthers and P. Smith. Cambridge University Press, pp. 200–220.

Heal, J. (2005). "Joint Attention and Understanding the Mind". In: *Joint Attention: Communication and Other Minds*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199245635.001.0001.

Heller, D., K. S. Gorman, and M. K. Tanenhaus (2012). "To Name or to Describe: Shared Knowledge Affects Referential Form". In: *Topics in Cognitive Science* 4.2, pp. 290–305. DOI: 10.1111/j.1756-8765.2012.01182.x.

Heller, D., D. Grodner, and M. K. Tanenhaus (2008). "The Role of Perspective in Identifying Domains of Reference". In: p. 6.

Helming, K. A., M. Sheskin, C. Chevalier, and T. C. Scott-Phillips (2015). "High Level Recursive Mindreading in Children". In: *Preprint*.

Helming, K. A., B. Strickland, and P. Jacob (2016). "Solving the Puzzle about Early Belief-Ascription". In: *Mind & Language* 31.4, pp. 438–469. DOI: 10.1111/mila.12114.

Henrich, J., S. J. Heine, and A. Norenzayan (2010). "The Weirdest People in the World?" In: *Behavioral and Brain Sciences* 33.2-3, pp. 61–83. DOI: 10.1017/S0140525X0999152X.

Heyes, C. (2014a). "False Belief in Infancy: A Fresh Look". In: *Developmental Science* 17.5, pp. 647–659. DOI: 10.1111/desc.12148.

— (2014b). "Submentalizing: I Am Not Really Reading Your Mind". In: *Perspectives on Psychological Science* 9.2, pp. 131–143. DOI: 10.1177/1745691613518076.

— (2015). "Animal Mindreading: What's the Problem?" In: *Psychonomic Bulletin & Review* 22.2, pp. 313–327. DOI: 10.3758/s13423-014-0704-4.

— (2018). "Enquire within: Cultural Evolution and Cognitive Science". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1743, p. 20170051. DOI: 10.1098/rstb.2017.0051.

Heyes, C. and C. D. Frith (2014). "The Cultural Evolution of Mind Reading". In: *Science* 344.6190, pp. 1243091–1243091. DOI: 10.1126/science.1243091.

Higginson, A. D. and M. R. Munafò (2016). "Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions". In: *PLOS Biology* 14.11, e2000995. DOI: 10.1371/journal.pbio.2000995.

Hobson, R. P., R. M. García-Pérez, and A. Lee (2010). "Person-Centred (Deictic) Expressions and Autism". In: *Journal of Autism and Developmental Disorders* 40.4, pp. 403–415. DOI: 10.1007/s10803-009-0882-5.

Hockett, C. F. (1960). "The Origin of Speech". In: *Scientific American*.

Hostetter, A. B., M. Cantero, and W. D. Hopkins (2001). "Differential Use of Vocal and Gestural Communication by Chimpanzees (Pan Troglodytes) in Response to the Attentional Status of a Human (Homo Sapiens)." In: *Journal of Comparative Psychology* 115.4, pp. 337–343. DOI: 10.1037/0735-7036.115.4.337.

Howlin, P. (2003). "Outcome in High-Functioning Adults with Autism With and Without Early Language Delays: Implications for the Differentiation Between Autism and Asperger Syndrome". In: *Journal of Autism and Developmental Disorders* 33.1, pp. 3–13. DOI: 10.1023/A:1022270118899.

Hudson Kam, C. L. and E. L. Newport (2005). "Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change". In: *Language Learning and Development* 1.2, pp. 151–195. DOI: 10.1080/15475441.2005.9684215.

Hughes, C., R. T. Devine, and Z. Wang (2018). "Does Parental Mind-Mindedness Account for Cross-Cultural Differences in Preschoolers' Theory of Mind?" In: *Child Development* 89.4, pp. 1296–1310. DOI: 10.1111/cdev.12746.

Hurford, J. R. (1989). "Biological Evolution of the Saussurean Sign as a Component of the Language Acquisition Device". In: *Lingua* 77.2, pp. 187–222. DOI: 10.1016/0024-3841(89)90015-6.

Ioannidis, J. P. A. (2005). "Why Most Published Research Findings Are False". In: *PLoS Medicine* 2.8, p. 6.

— (2008). "Why Most Discovered True Associations Are Inflated". In: *Epidemiology* 19.5, pp. 640–648. DOI: 10.1097/EDE.0b013e31818131e7.

— (2012). "Why Science Is Not Necessarily Self-Correcting". In: *Perspectives on Psychological Science* 7.6, pp. 645–654. DOI: 10.1177/1745691612464056.

Jakobsen, K. V., J. E. Frick, and E. A. Simpson (2013). "Look Here! The Development of Attentional Orienting to Symbolic Cues". In: *Journal of Cognition and Development* 14.2, pp. 229–249. DOI: 10.1080/15248372.2012.666772.

John, Leslie K, Loewenstein, George, and Prelec, Drazen (2012). "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling". In: *Psychological Science* 23.5, pp. 524–532.

Kalish, M. L., T. L. Griffiths, and S. Lewandowsky (2007). "Iterated Learning: Intergenerational Knowledge Transmission Reveals Inductive Biases". In: *Psychonomic Bulletin & Review* 14.2, pp. 288–294. DOI: 10.3758/BF03194066.

Kaminski, J., J. Call, and M. Tomasello (2004). "Body Orientation and Face Orientation: Two Factors Controlling Apes? Begging Behavior from Humans". In: *Animal Cognition* 7.4, pp. 216–223. DOI: 10.1007/s10071-004-0214-2.

Kazak, S., G. M. Collis, and V. Lewis (1997). "Can Young People with Autism Refer to Knowledge States? Evidence from Their Understanding of "Know" and "Guess"". In: *Journal of Child Psychology and Psychiatry* 38.8, pp. 1001–1009. DOI: 10.1111/j.1469-7610.1997.tb01617.x.

Kelley, E., J. J. Paul, D. Fein, and L. R. Naigles (2006). "Residual Language Deficits in Optimal Outcome Children with a History of Autism". In: *Journal of Autism and Developmental Disorders* 36.6, pp. 807–828. DOI: 10.1007/s10803-006-0111-4.

Kerr, N. L. (1998). "HARKing: Hypothesizing after the Results Are Known". In: *Personality and Social Psychology Review* 2.3, pp. 196–217.

Keysar, B. (2007). "Communication and Miscommunication: The Role of Egocentric Processes". In: *Intercultural Pragmatics* 4.1. DOI: 10.1515/IP.2007.004.

Keysar, B., D. J. Barr, J. A. Balin, and J. S. Brauner (2000). "Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension". In: *Psychological Science* 11.1, pp. 32–38. DOI: 10.1111/1467-9280.00211.

Keysar, B., S. Lin, and D. J. Barr (2003). "Limits on Theory of Mind Use in Adults". In: *Cognition* 89.1, pp. 25–41. DOI: 10.1016/S0010-0277(03)00064-7.

Kidwell, M. C. et al. (2016). "Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency". In: *PLOS Biology* 14.5. Ed. by M. R. Macleod, e1002456. DOI: 10.1371/journal.pbio.1002456.

Kinderman, P., R. I. M. Dunbar, and R. P. Bentall (1998). "Theory-of-Mind Deficits and Causal Attributions". In: *British Journal of Psychology* 89.2, pp. 191–204. DOI: 10.1111/j.2044-8295.1998.tb02680.x.

Kirby, S. (2000). "Syntax without Natural Selection: How Compositionality Emerges from Vocabulary in a Population of Learners". In: *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*. Cambridge University Press.

— (2001). "Spontaneous Evolution of Linguistic Structure-an Iterated Learning Model of the Emergence of Regularity and Irregularity". In: *IEEE Transactions on Evolutionary Computation* 5.2, pp. 102–110. DOI: 10.1109/4235.918430.

— (2002). "Learning, Bottlenecks and the Evolution of Recursive Syntax". In: *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

Kirby, S., H. Cornish, and K. Smith (2008). "Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686. DOI: 10.1073/pnas.0707835105.

Kirby, S., M. Dowman, and T. L. Griffiths (2007). "Innateness and Culture in the Evolution of Language". In: *Proceedings of the National Academy of Sciences* 104.12, pp. 5241–5245.

Kirby, S., T. L. Griffiths, and K. Smith (2014). "Iterated Learning and the Evolution of Language". In: *Current Opinion in Neurobiology* 28, pp. 108–114. DOI: 10.1016/j.conb.2014.07.014.

Kirby, S. and J. R. Hurford (2002). "The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model". In: *Simulating the Evolution of Language*. Springer Science & Business Media.

Kirby, S., M. Tamariz, H. Cornish, and K. Smith (2015). "Compression and Communication in the Cultural Evolution of Linguistic Structure". In: *Cognition* 141, pp. 87–102. DOI: 10.1016/j.cognition.2015.03.016.

Klein, R. A. et al. (2014). "Investigating Variation in Replicability: A "Many Labs" Replication Project". In: *Social Psychology* 45.3, pp. 142–152. DOI: 10.1027/1864-9335/a000178.

Kovacs, A. M., E. Teglas, and A. D. Endress (2010). "The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults". In: *Science* 330.6012, pp. 1830–1834. DOI: 10.1126/science.1190792.

Kristen, S., C. Thoermer, T. Hofer, G. Aschersleben, and B. Sodian (2006). "Scaling of "Theory of Mind" Tasks". In: *Journal of Developmental Psychology and Educational Psychology* 38, pp. 186–195.

Krupenye, C. and J. Call (2019). "Theory of Mind in Animals: Current and Future Directions". In: *Wiley Interdisciplinary Reviews: Cognitive Science*, e1503. DOI: 10.1002/wcs.1503.

Krupenye, C., F. Kano, S. Hirata, J. Call, and M. Tomasello (2016). "Great Apes Anticipate That Other Individuals Will Act According to False Beliefs". In: *Science* 354.6308, pp. 110–114. DOI: 10.1126/science.aaf8110.

Kuhlen, A. K. and S. E. Brennan (2010). "Anticipating Distracted Addressees: How Speakers' Expectations and Addressees' Feedback Influence Storytelling". In: *Discourse Processes* 47.7, pp. 567–587. DOI: 10.1080/01638530903441339.

Kulke, L., J. Johannsen, and H. Rakoczy (2019). "Why Can Some Implicit Theory of Mind Tasks Be Replicated and Others Cannot? A Test of Mentalizing versus Submentalizing Accounts". In: *PLOS ONE*.

Kulke, L. and H. Rakoczy (2018). "Implicit Theory of Mind – An Overview of Current Replications and Non-Replications". In: *Data in Brief* 16, pp. 101–104. DOI: 10.1016/j.dib.2017.11.016.

Kulke, L., B. von Duhn, D. Schneider, and H. Rakoczy (2018). "Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study". In: *Psychological Science* 29.6, pp. 888–900. DOI: 10.1177/0956797617747090.

Kuntoro, I. A., L. Saraswati, C. Peterson, and V. Slaughter (2013). "Micro-Cultural Influences on Theory of Mind Development: A Comparative Study of Middle-Class and *Pemulung* Children in Jakarta, Indonesia". In: *International Journal of Behavioral Development* 37.3, pp. 266–273. DOI: 10.1177/0165025413478258.

Ladefoged, P. (2006). *A Course in Phonetics*. Fifth. Thomson Wadsworth.

Lakens, D. et al. (2018). "Justify Your Alpha". In: *Nature Human Behaviour* 2.3, pp. 168–171. DOI: 10.1038/s41562-018-0311-x.

Langton, S. (2018). "I Don't See It Your Way: The Dot Perspective Task Does Not Gauge Spontaneous Perspective Taking". In: *Vision* 2.1, p. 6. DOI: 10.3390/vision2010006.

Lavelle, J. S. (2018). *The Social Mind: A Philosophical Introduction*. Routledge.

Leavens, D. A., A. B. Hostetter, M. J. Wesley, and W. D. Hopkins (2004). "Tactical Use of Unimodal and Bimodal Communication by Chimpanzees, Pan Troglodytes". In: *Animal Behaviour* 67.3, pp. 467–476. DOI: 10.1016/j.anbehav.2003.04.007.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Liebal, K., J. Call, M. Tomasello, and S. Pika (2004). "To Move or Not to Move: How Apes Adjust to the Attentional State of Others". In: *Interaction Studies* 5.2, pp. 199–219. DOI: 10.1075/is.5.2.03lie.

Lin, S., B. Keysar, and N. Epley (2010). "Reflexively Mindblind: Using Theory of Mind to Interpret Behavior Requires Effortful Attention". In: *Journal of Experimental Social Psychology* 46.3, pp. 551–556. DOI: 10.1016/j.jesp.2009.12.019.

Liszkowski, U., M. Carpenter, and M. Tomasello (2008). "Twelve-Month-Olds Communicate Helpfully and Appropriately for Knowledgeable and Ignorant Partners". In: *Cognition* 108.3, pp. 732–739. DOI: 10.1016/j.cognition.2008.06.013.

Lockridge, C. B. and S. E. Brennan (2002). "Addressees' Needs Influence Speakers' Early Syntactic Choices". In: *Psychonomic Bulletin & Review* 9.3, pp. 550–557. DOI: 10.3758/BF03196312.

Lohmann, H. and M. Tomasello (2003). "The Role of Language in the Development of False Belief Understanding: A Training Study". In: *Child Development* 74.4, pp. 1130–1144.

Lupyan, G. and R. Dale (2010). "Language Structure Is Partly Determined by Social Structure". In: *PLoS ONE* 5.1. Ed. by D. O'Rourke, e8559. DOI: 10.1371/journal.pone.0008559.

Lyn, H., J. L. Russell, and W. D. Hopkins (2010). "The Impact of Environment on the Comprehension of Declarative Communication in Apes". In: *Psychological Science* 21.3, pp. 360–365. DOI: 10.1177/0956797610362218.

Lyness, C., B. Woll, R. Campbell, and V. Cardin (2013). "How Does Visual Language Affect Crossmodal Plasticity and Cochlear Implant Success?" In: *Neuroscience & Biobehavioral Reviews* 37.10, pp. 2621–2630. DOI: 10.1016/j.neubiorev.2013.08.011.

Majid, A., M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson (2004). "Can Language Restructure Cognition? The Case for Space". In: *Trends in Cognitive Sciences* 8.3, pp. 108–114. DOI: 10.1016/j.tics.2004.01.003.

Makel, M. C., J. A. Plucker, and B. Hegarty (2012). "Replications in Psychology Research: How Often Do They Really Occur?" In: *Perspectives on Psychological Science* 7.6, pp. 537–542. DOI: 10.1177/1745691612460688.

Markman, E. M. and G. F. Wachtel (1988). "Children's Use of Mutual Exclusivity to Constrain the Meanings of Words". In: *Cognitive Psychology* 20.2, pp. 121–157. DOI: 10.1016/0010-0285(88)90017-5.

Marotta, A., J. Lupiáñez, D. Martella, and M. Casagrande (2012). "Eye Gaze versus Arrows as Spatial Cues: Two Qualitatively Different Modes of Attentional Selection." In: *Journal of Experimental Psychology: Human Perception and Performance* 38.2, pp. 326–335. DOI: 10.1037/a0023959.

Marshall, J., A. Gollwitzer, and L. R. Santos (2018). "Does Altercentric Interference Rely on Mentalizing?: Results from Two Level-1 Perspective-Taking Tasks". In: *PLOS ONE* 13.3. Ed. by S. Gilbert, e0194101. DOI: 10.1371/journal.pone.0194101.

Marszalek, J. M., C. Barber, J. Kohlhart, and B. H. Cooper (2011). "Sample Size in Psychological Research over the Past 30 Years". In: *Perceptual and Motor Skills* 112.2, pp. 331–348. DOI: 10.2466/03.11.PMS.112.2.331-348.

Marticorena, D. C., A. M. Ruiz, C. Mukerji, A. Goddu, and L. R. Santos (2011). "Monkeys Represent Others' Knowledge but Not Their Beliefs: Monkeys Represent Knowledge but Not Beliefs". In: *Developmental Science* 14.6, pp. 1406–1416. DOI: 10.1111/j.1467-7687.2011.01085.x.

Martin, A. and L. R. Santos (2016). "What Cognitive Representations Support Primate Theory of Mind?" In: *Trends in Cognitive Sciences* 20.5, pp. 375–382. DOI: 10.1016/j.tics.2016.03.005.

Maxwell, S. E. (2004). "The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies." In: *Psychological Methods* 9.2, pp. 147–163. DOI: 10.1037/1082-989X.9.2.147.

Mayer, A. and B. E. Träuble (2013). "Synchrony in the Onset of Mental State Understanding across Cultures? A Study among Children in Samoa". In: *International Journal of Behavioral Development* 37.1, pp. 21–28. DOI: 10.1177/0165025412454030.

Maylor, E. A. (1985). "Facilitatory and Inhibitory Components of Orienting in Visual Space". In: *Eleventh International Symposium on Attention and Performance*.

Maynard Smith, J. and D. G. C. Harper (2003). *Animal Signals*. Oxford University Press.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2019). "Abandon Statistical Significance". In: *The American Statistician* 73.sup1, pp. 235–245. DOI: 10.1080/00031305.2018.1527253.

Meristo, M., E. Hjelmquist, and G. Morgan (2011). "How Access to Language Affects Theory of Mind in Deaf Children". In: *Access to Language and Cognitive Development*. Ed. by M. Siegal and L. Surian. Oxford University Press, pp. 44–61. DOI: 10.1093/acprof:oso/9780199592722.003.0003.

Mesoudi, A. and A. Whiten (2008). "The Multiple Roles of Cultural Transmission Experiments in Understanding Human Cultural Evolution". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1509, pp. 3489–3501. DOI: 10.1098/rstb.2008.0129.

Mesoudi, A., A. Whiten, and R. I. M. Dunbar (2006). "A Bias for Social Information in Human Cultural Transmission". In: *British Journal of Psychology* 97.3, pp. 405–423. DOI: 10.1348/000712605X85871.

Michael, J., T. Wolf, C. Letesson, S. Butterfill, J. Skewes, and J. Hohwy (2018). "Seeing It Both Ways: Using a Double-Cuing Task to Investigate the Role of Spatial Cuing in Level-1 Visual Perspective-Taking." In: *Journal of Experimental Psychology: Human Perception and Performance* 44.5, pp. 693–702. DOI: 10.1037/xhp0000486.

Moeller, M. P. and B. Schick (2006). "Relations Between Maternal Input and Theory of Mind Understanding in Deaf Children". In: *Child Development* 77.3, pp. 751–766. DOI: 10.1111/j.1467-8624.2006.00901.x.

Moore, R. (2014). "Ontogenetic Constraints on Grice's Theory of Communication". In: *Trends in Language Acquisition Research*. Ed. by D. Matthews. Vol. 10. Amsterdam: John Benjamins Publishing Company, pp. 87–104. DOI: 10.1075/tilar.10.06moo.

Moore, R. (2016a). "Gricean Communication and Cognitive Development". In: *The Philosophical Quarterly*, pqw049. DOI: 10.1093/pq/pqw049.

— (2016b). "Meaning and Ostension in Great Ape Gestural Communication". In: *Animal Cognition* 19.1, pp. 223–231. DOI: 10.1007/s10071-015-0905-x.

Moors, A. and J. De Houwer (2006). "Automaticity: A Theoretical and Conceptual Analysis." In: *Psychological Bulletin* 132.2, pp. 297–326. DOI: 10.1037/0033-2909.132.2.297.

Munafò, M. R. et al. (2017). "A Manifesto for Reproducible Science". In: *Nature Human Behaviour* 1.1, p. 0021. DOI: 10.1038/s41562-016-0021.

Nadig, A. S. and J. C. Sedivy (2002). "Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution". In: *Psychological Science* 13.4, p. 8.

Naito, M. and K. Koyama (2006). "The Development of False-Belief Understanding in Japanese Children: Delay and Difference?" In: *International Journal of Behavioral Development* 30.4, pp. 290–304. DOI: 10.1177/0165025406063622.

Nielsen, M. K., L. Slade, J. P. Levy, and A. Holmes (2015). "Inclined to See It Your Way: Do Altercentric Intrusion Effects in Visual Perspective Taking Reflect an Intrinsically Social Process?" In: *Quarterly Journal of Experimental Psychology* 68.10, pp. 1931–1951. DOI: 10.1080/17470218.2015.1023206.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor (2018). "The Preregistration Revolution". In: *Proceedings of the National Academy of Sciences* 115.11, pp. 2600–2606. DOI: 10.1073/pnas.1708274114.

Nosek, B. A. and D. Lakens (2014). "Registered Reports: A Method to Increase the Credibility of Published Results". In: *Social Psychology* 45.3, pp. 137–141. DOI: 10.1027/1864-9335/a000192.

Nosek, B. A. and S. Lindsay (2018). "Preregistration Becoming the Norm in Psychological Science". In: *APS Observer*.

Nosek, B. A. et al. (2015). "Promoting an Open Research Culture". In: *Science* 348.6242, pp. 1422–1425. DOI: 10.1126/science.aab2374.

Novogrodsky, R. (2013). "Subject Pronoun Use by Children with Autism Spectrum Disorders (ASD)". In: *Clinical Linguistics & Phonetics* 27.2, pp. 85–93. DOI: 10.3109/02699206.2012.742567.

O'Grady, C., C. Kliesch, K. Smith, and T. C. Scott-Phillips (2015). "The Ease and Extent of Recursive Mindreading, across Implicit and Explicit Tasks". In: *Evolution and Human Behavior* 36.4, pp. 313–322. DOI: 10.1016/j.evolhumbehav.2015.01.004.

O'Grady, C., T. C. Scott-Phillips, J. S. Lavelle, and K. Smith (2017). "The Dot Perspective Task Revisited: Evidence for Directional Effects". In: *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

O'Grady, C. and K. Smith (2018). "Models of Language Evolution". In: *The Oxford Handbook of Psycholinguistics*. Oxford University Press, p. 17.

O'Reilly, K., C. C. Peterson, and H. M. Wellman (2014). "Sarcasm and Advanced Theory of Mind Understanding in Children and Adults with Prelingual Deafness." In: *Developmental Psychology* 50.7, pp. 1862–1877. DOI: 10.1037/a0036654.

Olson, D. R. (1988). "On the Origins of Beliefs and Other Intentional States in Children". In: *Developing theories of mind* 6, pp. 414–426.

Onishi, K. H. and R. Baillargeon (2005). "Do 15-Month-Old Infants Understand False Beliefs?" In: *Science* 308.5719, pp. 255–258. DOI: 10.1126/science.1107621.

Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science". In: *Science* 349.6251, aac4716–aac4716. DOI: 10.1126/science.aac4716.

Orben, A. and A. K. Przybylski (2019). "The Association between Adolescent Well-Being and Digital Technology Use". In: *Nature Human Behaviour* 3.2, pp. 173–182. DOI: 10.1038/s41562-018-0506-1.

Osherovich, L. (2011). "Hedging against Academic Risk". In: *Science-Business eXchange* 4.15, pp. 416–416. DOI: 10.1038/scibx.2011.416.

Ostojić, L., E. W. Legg, R. C. Shaw, L. G. Cheke, M. Mendl, and N. S. Clayton (2014). "Can Male Eurasian Jays Disengage from Their Own Current Desire to Feed the Female What She Wants?" In: *Biology Letters* 10.3, p. 20140042. DOI: 10.1098/rsbl.2014.0042.

Ostojić, L., R. C. Shaw, L. G. Cheke, and N. S. Clayton (2013). "Evidence Suggesting That Desire-State Attribution May Govern Food Sharing in Eurasian Jays". In: *Proceedings of the National Academy of Sciences* 110.10, pp. 4123–4128. DOI: 10.1073/pnas.1209926110.

Oudeyer, P.-Y. (2005a). *Self-Organization in the Evolution of Speech*. Oxford University Press.

— (2005b). "The Self-Organization of Speech Sounds". In: *arXiv:cs/0502086*. arXiv: cs/0502086.

Oxford English Dictionary (2016). *Dictionary Facts*.

Parish-Morris, J., K. Hirsh-Pasek, E. A. Hennon, R. M. Golinkoff, and H. Tager-Flusberg (2007). "Children with Autism Illuminate the Role of Social Intention in Word Learning". In: *Child Development* 78.4, pp. 1265–1287.

Pashler, H. and C. R. Harris (2012). "Is the Replicability Crisis Overblown? Three Arguments Examined". In: *Perspectives on Psychological Science* 7.6, pp. 531–536.

Peacocke, C. (2005). "Joint Attention: Its Nature, Reflexivity, and Relation to Common Knowledge". In: *Joint Attention: Communication and Other Minds*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199245635.001.0001.

Peirce, J. (2010). "PsychoPy – Psychology Software for Python". In: p. 289.

Perfors, A. and D. J. Navarro (2014). "Language Evolution Can Be Shaped by the Structure of the World". In: *Cognitive Science* 38.4, pp. 775–793. DOI: 10.1111/cogs.12102.

Perner, J., S. R. Leekam, and H. Wimmer (1987). "Three-Year-Olds' Difficulty with False Belief: The Case for a Conceptual Deficit". In: *British Journal of Developmental Psychology* 5.2, pp. 125–137. DOI: 10.1111/j.2044-835X.1987.tb01048.x.

Peterson, C. C. and M. Siegal (2000). "Insights into Theory of Mind from Deafness and Autism". In: *Mind and Language* 15.1, pp. 123–145. DOI: 10.1111/1468-0017.00126.

Peterson, C. C., H. M. Wellman, and D. Liu (2005). "Steps in Theory-of-Mind Development for Children With Deafness or Autism". In: *Child Development* 76.2, pp. 502–517. DOI: 10.1111/j.1467-8624.2005.00859.x.

Phillips, J., D. C. Ong, A. D. R. Surtees, Y. Xin, S. Williams, R. Saxe, and M. C. Frank (2015). "A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010)". In: *Psychological Science* 26.9, pp. 1353–1367. DOI: 10.1177/0956797614558717.

Pickering, M. J. and S. Garrod (2004). "Toward a Mechanistic Psychology of Dialogue". In: *Behavioral and Brain Sciences* 27.02. DOI: 10.1017/S0140525X04000056.

Pickles, A., D. K. Anderson, and C. Lord (2014). "Heterogeneity and Plasticity in the Development of Language: A 17-Year Follow-up of Children Referred Early for Possible Autism". In: *Journal of Child Psychology and Psychiatry* 55.12, pp. 1354–1362. DOI: 10.1111/jcpp.12269.

Pinker, S. and P. Bloom (1990). "Natural Language and Natural Selection". In: *Behavioral and Brain Sciences* 13.4, pp. 707–784.

Posner, M. I. (1980). "Orienting of Attention". In: *Quarterly Journal of Experimental Psychology* 32.1, pp. 3–25.

Posner, M. I. and Y. Cohen (1984). "Components of Visual Orienting". In: *Attention and performance: Control of language processes* 32, pp. 531–556.

Povinelli, D. J., T. J. Eddy, R. P. Hobson, and M. Tomasello (1996). "What Young Chimpanzees Know about Seeing". In: *Monographs of the Society for Research in Child Development* 61.3, p. i. DOI: 10.2307/1166159.

Povinelli, D. J., L. A. Theall, J. E. Reaux, and S. Dunphy-Lelii (2003). "Chimpanzees Spontaneously Alter the Location of Their Gestures to Match the Attentional Orientation of Others". In: *Animal Behaviour* 66.1, pp. 71–79. DOI: 10.1006/anbe.2003.2195.

Powell, L. J., K. Hobbs, A. Bardis, S. Carey, and R. Saxe (2018). "Replications of Implicit Theory of Mind Tasks with Varying Representational Demands". In: *Cognitive Development* 46, pp. 40–50. DOI: 10.1016/j.cogdev.2017.10.004.

Preissler, M. A. and S. Carey (2005). "The Role of Inferences about Referential Intent in Word Learning: Evidence from Autism". In: *Cognition* 97.1, B13–B23. DOI: 10.1016/j.cognition.2005.01.008.

Premack, P. and G. Woodruff (1978). "Does the Chimpanzee Have a Theory of Mind?" In: *Behavioral and Brain Sciences*.

Prinz, F., T. Schlange, and K. Asadullah (2011). "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" In: *Nature Reviews Drug Discovery* 10.9, pp. 712–712. DOI: 10.1038/nrd3439-c1.

Pyers, J. E. and A. Senghas (2009). "Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language". In: *Psychological Science* 20.7, pp. 805–812. DOI: 10.1111/j.1467-9280.2009.02377.x.

Qureshi, A. W., I. A. Apperly, and D. Samson (2010). "Executive Function Is Necessary for Perspective Selection, Not Level-1 Visual Perspective Calculation: Evidence from a Dual-Task Study of Adults". In: *Cognition* 117.2, pp. 230–236. DOI: 10.1016/j.cognition.2010.08.003.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*.

Rakoczy, H. (2017). "In Defense of a Developmental Dogma: Children Acquire Propositional Attitude Folk Psychology around Age 4". In: *Synthese* 194.3, pp. 689–707. DOI: 10.1007/s11229-015-0860-8.

Ramscar, M., D. Yarlett, M. Dye, K. Denny, and K. Thorpe (2010). "The Effects of Feature-Label-Order and Their Implications for Symbolic Learning". In: *Cognitive Science* 34.6, pp. 909–957. DOI: 10.1111/j.1551-6709.2009.01092.x.

Reali, F. and T. L. Griffiths (2009). "The Evolution of Frequency Distributions: Relating Regularization to Inductive Biases through Iterated Learning". In: *Cognition* 111.3, pp. 317–328. DOI: 10.1016/j.cognition.2009.02.012.

Ristic, J., C. K. Friesen, and A. Kingstone (2002). "Are Eyes Special? It Depends on How You Look at It". In: *Psychonomic Bulletin & Review* 9.3, pp. 507–513. DOI: 10.3758/BF03196306.

Ristic, J. and A. Kingstone (2005). "Taking Control of Reflexive Social Attention". In: *Cognition* 94.3, B55–B65. DOI: 10.1016/j.cognition.2004.04.005.

Robbins, J. and A. Rumsey (2008). "Introduction: Cultural and Linguistic Anthropology and the Opacity of Other Minds". In: *Anthropological Quarterly* 81.2, pp. 407–420. DOI: 10.1353/anq.0.0005.

Roberts, G. and B. Galantucci (2012). "The Emergence of Duality of Patterning: Insights from the Laboratory". In: *Language and Cognition* 4.04, pp. 297–318. DOI: 10.1515/langcog-2012-0017.

Rosenthal, R. (1979). "The "File Drawer Problem" and Tolerance for Null Results". In: *Psychological Bulletin* 86.3, pp. 638–641.

Rubio-Fernández, P. (2017). "The Director Task: A Test of Theory-of-Mind Use or Selective Attention?" In: *Psychonomic Bulletin & Review* 24.4, pp. 1121–1128. DOI: 10.3758/s13423-016-1190-7.

Rubio-Fernández, P. and B. Geurts (2013). "How to Pass the False-Belief Task Before Your Fourth Birthday". In: *Psychological Science* 24.1, pp. 27–33. DOI: 10.1177/0956797612447819.

Russell, J. L., H. Lyn, J. A. Schaeffer, and W. D. Hopkins (2011). "The Role of Socio-Communicative Rearing Environments in the Development of Social and Physical Cognition in Apes: Development of Social and Physical Cognition in Apes". In: *Developmental Science* 14.6, pp. 1459–1470. DOI: 10.1111/j.1467-7687.2011.01090.x.

Rutherford, M. D. (2004). "The Effect of Social Role on Theory of Mind Reasoning". In: *British Journal of Psychology* 95.1, pp. 91–103. DOI: 10.1348/000712604322779488.

Samson, D., I. A. Apperly, J. J. Braithwaite, B. J. Andrews, and S. E. Bodley Scott (2010). "Seeing It Their Way: Evidence for Rapid and Involuntary Computation of What Other People See." In: *Journal of Experimental Psychology: Human Perception and Performance* 36.5, pp. 1255–1266. DOI: 10.1037/a0018729.

Sandler, W., M. Aronoff, I. Meir, and C. Padden (2011). "The Gradual Emergence of Phonological Form in a New Language". In: *Natural Language & Linguistic Theory* 29.2, pp. 503–543. DOI: 10.1007/s11049-011-9128-2.

Santiesteban, I., C. Catmur, S. C. Hopkins, G. Bird, and C. Heyes (2014). "Avatars and Arrows: Implicit Mentalizing or Domain-General Processing?" In: *Journal of Experimental Psychology: Human Perception and Performance* 40.3, pp. 929–937. DOI: 10.1037/a0035175.

Schel, A. M., Z. Machanda, S. W. Townsend, K. Zuberbühler, and K. E. Slocombe (2013a). "Chimpanzee Food Calls Are Directed at Specific Individuals". In: *Animal Behaviour* 86.5, pp. 955–965. DOI: 10.1016/j.anbehav.2013.08.013.

Schel, A. M., S. W. Townsend, Z. Machanda, K. Zuberbühler, and K. E. Slocombe (2013b). "Chimpanzee Alarm Call Production Meets Key Criteria for Intentionality". In: *PLoS ONE* 8.10. Ed. by K. McComb, e76674. DOI: 10.1371/journal.pone.0076674.

Schick, B., P. de Villiers, J. de Villiers, and R. Hoffmeister (2007). "Language and Theory of Mind: A Study of Deaf Children". In: *Child Development* 78.2, pp. 376–396.

Schiffer, S. R. (1972). *Meaning*. Clarendon Press.

Schilbach, L., B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, and K. Vogeley (2013). "Toward a Second-Person Neuroscience". In: *Behavioral and Brain Sciences* 36.04, pp. 393–414. DOI: 10.1017/S0140525X12000660.

Schmidt, F. L. (1996). "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers". In: *Psychological Methods* 1.2, p. 115.

Schneider, D., A. Grigutsch, M. Schurz, R. Zäske, and S. R. Schweinberger (2018). "Group Membership and the Effects on Visual Perspective Taking (Preprint)". In: DOI: 10.31234/osf.io/wnrk6.

Schneider, D., R. Lam, A. P. Bayliss, and P. E. Dux (2012). "Cognitive Load Disrupts Implicit Theory-of-Mind Processing". In: *Psychological Science* 23.8, pp. 842–847. DOI: 10.1177/0956797612439070.

Schneider, D., Z. E. Nott, and P. E. Dux (2014). "Task Instructions and Implicit Theory of Mind". In: *Cognition* 133.1, pp. 43–47. DOI: 10.1016/j.cognition.2014.05.016.

Schneider, D., V. P. Slaughter, and P. E. Dux (2017). "Current Evidence for Automatic Theory of Mind Processing in Adults". In: *Cognition* 162, pp. 27–31. DOI: 10.1016/j.cognition.2017.01.018.

Schouwstra, M. and H. de Swart (2014). "The Semantic Origins of Word Order". In: *Cognition* 131.3, pp. 431–436. DOI: 10.1016/j.cognition.2014.03.004.

Schouwstra, M., K. Smith, and S. Kirby (2016). "From Natural Order to Convention in Silent Gesture". In: *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Schurz, M., M. Kronbichler, S. Weissengruber, A. D. R. Surtees, D. Samson, and J. Perner (2015). "Clarifying the Role of Theory of Mind Areas during Visual Perspective Taking: Issues of

Spontaneity and Domain-Specificity". In: *NeuroImage* 117, pp. 386–396. DOI: 10.1016/j.neuroimage.2015.04.031.

Schuwerk, T., B. Priewasser, B. Sodian, and J. Perner (2018). "The Robustness and Generalizability of Findings on Spontaneous False Belief Sensitivity: A Replication Attempt". In: *Royal Society Open Science* 5.5, p. 172273. DOI: 10.1098/rsos.172273.

Schwarzkopf, S., L. Schilbach, K. Vogeley, and B. Timmermans (2014). ""Making It Explicit" Makes a Difference: Evidence for a Dissociation of Spontaneous and Intentional Level 1 Perspective Taking in High-Functioning Autism". In: *Cognition* 131.3, pp. 345–354. DOI: 10.1016/j.cognition.2014.02.003.

Scott-Phillips, T. C. (2015). *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make It Special*. OCLC: 897806879. Basingstoke: Palgrave Macmillan.

— (2016). "Meaning in Great Ape Communication: Summarising the Debate". In: *Animal Cognition* 19.1, pp. 233–238. DOI: 10.1007/s10071-015-0936-3.

Scott-Phillips, T. C., R. A. Blythe, A. Gardner, and S. A. West (2012). "How Do Communication Systems Emerge?" In: *Proceedings of the Royal Society B: Biological Sciences* 279.1735, pp. 1943–1949. DOI: 10.1098/rspb.2011.2181.

Scott-Phillips, T. C., S. Kirby, and G. R. S. Ritchie (2009). "Signalling Signalhood and the Emergence of Communication". In: *Cognition*, p. 9.

Scott, R. M. and R. Baillargeon (2009). "Which Penguin Is This? Attributing False Beliefs About Object Identity at 18 Months". In: *Child Development* 80.4, pp. 1172–1196. DOI: 10.1111/j.1467-8624.2009.01324.x.

— (2017). "Early False-Belief Understanding". In: *Trends in Cognitive Sciences* 21.4, pp. 237–249. DOI: 10.1016/j.tics.2017.01.012.

Scott, R. M., E. Roby, and M. Smith (2016). "False-Belief Understanding in the First Years of Life". In: *Routledge Handbook of the Philosophy of the Social Mind*. Ed. by J. Kiverstein, pp. 152–171.

Senju, A., V. Southgate, S. J. White, and U. Frith (2009). "Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome". In: *Science* 325.5942, pp. 883–885. DOI: 10.1126/science.1176170.

Seyfarth, R. M., D. L. Cheney, and P. Marler (1980). "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication". In: *Science* 210.4471, pp. 801–803.

Shadish, W. R., M. Doherty, and L. M. Montgomery (1989). "How Many Studies Are in the File Drawer? An Estimate from the Family/Marital Psychotherapy Literature". In: *Clinical Psychology Review* 9.5, pp. 589–603. DOI: 10.1016/0272-7358(89)90013-5.

Shahaeian, A., C. C. Peterson, V. Slaughter, and H. M. Wellman (2011). "Culture and the Sequence of Steps in Theory of Mind Development." In: *Developmental Psychology* 47.5, pp. 1239–1247. DOI: 10.1037/a0023899.

Shiffrin, R. M. and W. Schneider (1977). "Controlled and Automatic Human Information Processing: II. Perceptual Learning, Automatic Attending, and a General Theory". In: *Psychological Review* 84.2.

Silberzahn, R. et al. (2018). "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results". In: *Advances in Methods and Practices in Psychological Science* 1.3, pp. 337–356. DOI: 10.1177/2515245917747646.

Silvey, C., S. Kirby, and K. Smith (2015). "Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions". In: *Cognitive Science* 39.1, pp. 212–226. DOI: 10.1111/cogs.12150.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11, pp. 1359–1366. DOI: 10.1177/0956797611417632.

— (2018). "False-Positive Citations". In: *Perspectives on Psychological Science* 13.2, pp. 255–259. DOI: 10.1177/1745691617698146.

Simonsohn, U. (2015). "Small Telescopes: Detectability and the Evaluation of Replication Results". In: *Psychological Science* 26.5, pp. 559–569. DOI: 10.1177/0956797614567341.

Singmann, H., B. Bolker, J. Westfall, and F. Aust (2017). "Afex: Analysis of Factorial Experiments". In: *R package version 0.17-8*.

Siposova, B. and M. Carpenter (2019). "A New Look at Joint Attention and Common Knowledge". In: *Cognition* 189, pp. 260–274. DOI: 10.1016/j.cognition.2019.03.019.

Slaughter, V. and D. Perez-Zapata (2014). "Cultural Variations in the Development of Mind Reading". In: *Child Development Perspectives* 8.4, pp. 237–241. DOI: 10.1111/cdep.12091.

Slaughter, V. and C. C. Peterson (2011). "How Conversational Input Shapes Theory of Mind Development in Infancy and Early Childhood". In: *Access to Language and Cognitive Development*. Ed. by M. Siegal and L. Surian. Oxford University Press, pp. 4–22. DOI: 10.1093/acprof:oso/9780199592722.003.0001.

Smith, K. and S. Kirby (2008). "Cultural Evolution: Implications for Understanding the Human Language Faculty and Its Evolution". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1509, pp. 3591–3603. DOI: 10.1098/rstb.2008.0145.

Smith, K. and E. Wonnacott (2010). "Eliminating Unpredictable Variation through Iterated Learning". In: *Cognition* 116.3, pp. 444–449. DOI: 10.1016/j.cognition.2010.06.004.

Song, H.-j. and R. Baillargeon (2008). "Infants' Reasoning about Others' False Perceptions." In: *Developmental Psychology* 44.6, pp. 1789–1795. DOI: 10.1037/a0013774.

Song, H.-j., K. H. Onishi, R. Baillargeon, and C. Fisher (2008). "Can an Agent's False Belief Be Corrected by an Appropriate Communication? Psychological Reasoning in 18-Month-Old Infants". In: *Cognition* 109.3, pp. 295–315. DOI: 10.1016/j.cognition.2008.08.008.

Southgate, V., A. Senju, and G. Csibra (2007). "Action Anticipation Through Attribution of False Belief by 2-Year-Olds". In: *Psychological Science* 18.7, pp. 587–592. DOI: 10.1111/j.1467-9280.2007.01944.x.

Sperber, D. (2000a). "Metarepresentations in an Evolutionary Perspective". In: *Metarepresentations: A Multidisciplinary Perspective*. Oxford University Press.

— ed. (2000b). *Metarepresentations: A Multidisciplinary Perspective*. Vancouver Studies in Cognitive Science v. 10. Oxford ; New York: Oxford University Press.

Sperber, D. and G. Origgi (2012). "A Pragmatic Perspective on the Evolution of Language". In: *Meaning and Relevance*. Cambridge University Press.

Sperber, D. and D. Wilson (1986). *Relevance: Communication and Cognition*. 2nd ed. Oxford ; Cambridge, MA: Blackwell Publishers.

— (2002). "Pragmatics, Modularity and Mind-Reading". In: p. 21.

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel (2016). "Increasing Transparency Through a Multiverse Analysis". In: *Perspectives on Psychological Science* 11.5, pp. 702–712. DOI: 10.1177/1745691616658637.

Steels, L. (1999). *The Talking Heads Experiment*. Laboratorium.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam (1995). "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa". In: *The American Statistician* 49.1, pp. 108–112.

Stiller, J. and R. I. M. Dunbar (2007). "Perspective-Taking and Memory Capacity Predict Social Network Size". In: *Social Networks* 29.1, pp. 93–104. DOI: 10.1016/j.socnet.2006.04.001.

Strawson, P. F. (1964). "Intention and Convention in Speech Acts". In: *The Philosophical Review* 73.4, p. 439. DOI: 10.2307/2183301.

Surian, L., S. Baron-Cohen, and H. Van der Lely (1996). "Are Children with Autism Deaf to Gricean Maxims?" In: *Cognitive Neuropsychiatry* 1.1, pp. 55–72. DOI: 10.1080/135468096396703.

Surian, L., S. Caldi, and D. Sperber (2007). "Attribution of Beliefs by 13-Month-Old Infants". In: *Psychological Science* 18.7, pp. 580–586. DOI: 10.1111/j.1467-9280.2007.01943.x.

Surtees, A. D. R. and I. A. Apperly (2012). "Egocentrism and Automatic Perspective Taking in Children and Adults: Egocentrism and Automatic Perspective Taking". In: *Child Development*, no–no. DOI: 10.1111/j.1467-8624.2011.01730.x.

Surtees, A. D. R., I. A. Apperly, and D. Samson (2013). "The Use of Embodied Self-Rotation for Visual and Spatial Perspective-Taking". In: *Frontiers in Human Neuroscience* 7. DOI: 10.3389/fnhum.2013.00698.

— (2016a). "I've Got Your Number: Spontaneous Perspective-Taking in an Interactive Task". In: *Cognition* 150, pp. 43–52. DOI: 10.1016/j.cognition.2016.01.014.

Surtees, A. D. R., S. A. Butterfill, and I. A. Apperly (2012). "Direct and Indirect Measures of Level-2 Perspective-Taking in Children and Adults: Level-2 Perspective-Taking". In: *British Journal of Developmental Psychology* 30.1, pp. 75–86. DOI: 10.1111/j.2044-835X.2011.02063.x.

Surtees, A. D. R., D. Samson, and I. A. Apperly (2016b). "Unintentional Perspective-Taking Calculates Whether Something Is Seen, but Not How It Is Seen". In: *Cognition* 148, pp. 97–105. DOI: 10.1016/j.cognition.2015.12.010.

Szucs, D. and J. P. A. Ioannidis (2017). "Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature". In: *PLOS Biology* 15.3. Ed. by E.-J. Wagenmakers, e2000797. DOI: 10.1371/journal.pbio.2000797.

Tager-Flusberg, H. (1992). "Autistic Children's Talk about Psychological States: Deficits in the Early Acquisition of a Theory of Mind". In: *Child Development*, p. 13.

Tager-Flusberg, H., R. Paul, and C. Lord (2013). "Language and Communication in Autism". In: *Handbook of Autism and Pervasive Developmental Disorders*. Ed. by F. R. Volkmar, R. Paul, A. Klin, and D. Cohen. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 335–364. DOI: 10.1002/9780470939345.ch12.

Taumoepeau, M. (2015). "From Talk to Thought: Strength of Ethnic Identity and Caregiver Mental State Talk Predict Social Understanding in Preschoolers". In: *Journal of Cross-Cultural Psychology* 46.9, pp. 1169–1190. DOI: 10.1177/0022022115604393.

Taumoepeau, M. and T. Ruffman (2006). "Mother and Infant Talk about Mental States Relates to Desire Language and Emotion Understanding". In: *Child Development* 77.2, pp. 465–481.

— (2008). "Stepping Stones to Others' Minds: Maternal Talk Relates to Child Mental State Language and Emotion Understanding at 15, 24, and 33 Months". In: *Child Development* 79.2, pp. 284–302. DOI: 10.1111/j.1467-8624.2007.01126.x.

Theisen-White, C., S. Kirby, and J. Oberlander (2011). "Integrating the Horizontal and Vertical Cultural Transmission of Novel Communication Systems". In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Theisen, C. A., C. A. Theisen, J. Oberlander, and S. Kirby (2009). "Systematicity and Arbitrariness in Novel Communication Systems". In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Thompson, B., S. Kirby, and K. Smith (2016). "Culture Shapes the Evolution of Cognition". In: *Proceedings of the National Academy of Sciences* 113.16, pp. 4530–4535.

Tomasello, M. (2000). "The Social-Pragmatic Theory of Word Learning". In: *Pragmatics* 10.4, pp. 401–413. DOI: 10.1075/prag.10.4.01tom.

— (2008). *Origins of Human Communication*. The Jean Nicod Lectures 2008. OCLC: ocn182779723. Cambridge, Mass: MIT Press.

— (2018). "How Children Come to Understand False Beliefs: A Shared Intentionality Account". In: *Proceedings of the National Academy of Sciences* 115.34, pp. 8491–8498. DOI: 10.1073/pnas.1804761115.

Tomasello, M. and M. Barton (1994). "Learning Words in Nonostensive Contexts". In: *Developmental Psychology* 30.5, pp. 639–650.

Tomasello, M., J. Call, K. Nagell, R. Olguin, and M. Carpenter (1994). "The Learning and Use of Gestural Signals by Young Chimpanzees: A Transgenerational Study". In: *Primates* 35.2, pp. 137–154.

Tomasello, M. and K. Haberl (2003). "Understanding Attention: 12- and 18-Month-Olds Know What Is New for Other Persons." In: *Developmental Psychology* 39.5, pp. 906–912. DOI: 10.1037/0012-1649.39.5.906.

Townsend, S. W. et al. (2017). "Exorcising Grice's Ghost: An Empirical Approach to Studying Intentional Communication in Animals: Intentional Communication in Animals". In: *Biological Reviews* 92.3, pp. 1427–1433. DOI: 10.1111/brv.12289.

Tversky, B. and B. M. Hard (2009). "Embodied and Disembodied Cognition: Spatial Perspective-Taking". In: *Cognition* 110.1, pp. 124–129. DOI: 10.1016/j.cognition.2008.10.008.

van 't Veer, A. E. and R. Giner-Sorolla (2016). "Pre-Registration in Social Psychology—A Discussion and Suggested Template". In: *Journal of Experimental Social Psychology* 67, pp. 2–12. DOI: 10.1016/j.jesp.2016.03.004.

van der Wel, R. P., N. Sebanz, and G. Knoblich (2014). "Do People Automatically Track Others' Beliefs? Evidence from a Continuous Measure". In: *Cognition* 130.1, pp. 128–133. DOI: 10.1016/j.cognition.2013.10.004.

Van Duijn, M. (2010). "Mind the Reader". PhD thesis. Leiden University.

Vankov, I., J. Bowers, and M. R. Munafò (2014). "Article Commentary: On the Persistence of Low Power in Psychological Science". In: *Quarterly Journal of Experimental Psychology* 67.5, pp. 1037–1040. DOI: 10.1080/17470218.2014.885986.

Verhoef, T., S. Kirby, and C. Padden (2011). "Cultural Emergence of Combinatorial Structure in an Artificial Whistled Language". In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Wagenmakers, E.-J., R. Wetzels, D. Borsboom, H. L. J. van der Maas, and R. A. Kievit (2012). "An Agenda for Purely Confirmatory Research". In: *Perspectives on Psychological Science* 7.6, pp. 632–638. DOI: 10.1177/1745691612463078.

Wedel, A. (2006). "Exemplar Models, Evolution and Language Change". In: *The Linguistic Review* 23.3. DOI: 10.1515/TLR.2006.010.

— (2012). "Lexical Contrast Maintenance and the Organization of Sublexical Contrast Systems". In: *Language and Cognition* 4.04, pp. 319–355. DOI: 10.1515/langcog-2012-0018.

Wegner, D. M., M. Ansfield, and D. Pilloff (1998). "The Putt and the Pendulum: Ironic Effects of the Mental Control of Action". In: *Psychological Science* 9, pp. 1996–1999.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199334919.001.0001.

Wellman, H. M., F. Fang, D. Liu, L. Zhu, and G. Liu (2006). "Scaling of Theory-of-Mind Understandings in Chinese Children". In: *Psychological Science* 17.12, pp. 1075–1081. DOI: 10.1111/j.1467-9280.2006.01830.x.

Wellman, H. M. and D. Liu (2004). "Scaling of Theory-of-Mind Tasks". In: *Child Development* 75.2, pp. 523–541. DOI: 10.1111/j.1467-8624.2004.00691.x.

Westra, E. (2017a). "Pragmatic Development and the False Belief Task". In: *Review of Philosophy and Psychology* 8.2, pp. 235–257. DOI: 10.1007/s13164-016-0320-5.

— (2017b). "Spontaneous Mindreading: A Problem for the Two-Systems Account". In: *Synthese* 194.11, pp. 4559–4581. DOI: 10.1007/s11229-016-1159-0.

Westra, E. (2017c). "The Architecture and Development of Mindreading: Beliefs, Perspectives and Character". PhD thesis. University of Maryland.

Westra, E. and P. Carruthers (2017). "Pragmatic Development Explains the Theory-of-Mind Scale". In: *Cognition* 158, pp. 165–176. DOI: 10.1016/j.cognition.2016.10.021.

Wetzels, R., D. Matzke, M. D. Lee, J. N. Rouder, G. J. Iverson, and E.-J. Wagenmakers (2011). "Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 *t* Tests". In: *Perspectives on Psychological Science* 6.3, pp. 291–298. DOI: 10.1177/1745691611406923.

Whelan, R. (2008). "Effective Analysis of Reaction Time Data". In: *The Psychological Record* 58.3, pp. 475–482. DOI: 10.1007/BF03395630.

Whiten, A., C. A. Caldwell, and A. Mesoudi (2016). "Cultural Diffusion in Humans and Other Animals". In: *Current Opinion in Psychology* 8, pp. 15–21. DOI: 10.1016/j.copsyc.2015.09.002.

Wicherts, J. M., M. Bakker, and D. Molenaar (2011). "Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results". In: *PLoS ONE* 6.11. Ed. by R. E. Tractenberg, e26828. DOI: 10.1371/journal.pone.0026828.

Wicherts, J. M., D. Borsboom, J. Kats, and D. Molenaar (2006). "The Poor Availability of Psychological Research Data for Reanalysis." In: *American Psychologist* 61.7, pp. 726–728. DOI: 10.1037/0003-066X.61.7.726.

Wiese, E., A. Wykowska, J. Zwickel, and H. J. Müller (2012). "I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others". In: *PLoS ONE* 7.9. Ed. by S. B. Hamed, e45391. DOI: 10.1371/journal.pone.0045391.

Wilby, M. (2010). "The Simplicity of Mutual Knowledge". In: *Philosophical Explorations* 13.2, pp. 83–100. DOI: 10.1080/13869791003759963.

Wilkinson, L. and Task Force on Statistical Inference (1999). "Statistical Methods in Psychology Journals". In: *American Psychologist*.

Wilson, C. J., A. Soranzo, and M. Bertamini (2017). "Attentional Interference Is Modulated by Salience Not Sentience". In: *Acta Psychologica* 178, pp. 56–65. DOI: 10.1016/j.actpsy.2017.05.010.

Wilson, C. (2003). "Experimental Investigation of Phonological Naturalness". In: *West Coast Conference on Formal Linguistics 22 (WCCFL22)*, pp. 101–114.

Wilson, D. (2013). *Beyond Speaker's Meaning*. DOI: 10.1037/e505052014-060.

Wimmer, H. and J. Perner (1983). "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception". In: *Cognition* 13, pp. 103–128.

Winters, J., S. Kirby, and K. Smith (2015). "Languages Adapt to Their Contextual Niche". In: *Language and Cognition* 7.03, pp. 415–449. DOI: 10.1017/langcog.2014.35.

Wodka, E. L., P. Mathy, and L. Kalb (2013). "Predictors of Phrase and Fluent Speech in Children With Autism and Severe Language Delay". In: *Pediatrics* 131.4, e1128–e1134. DOI: 10.1542/peds.2012-2221.

Wonnacott, E. and E. L. Newport (2005). "Novelty and Regularization: The Effect of Novel Instances on Rule Formation". In: *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*.

Wray, A. and G. W. Grace (2007). "The Consequences of Talking to Strangers: Evolutionary Corollaries of Socio-Cultural Influences on Linguistic Form". In: *Lingua* 117.3, pp. 543–578. DOI: 10.1016/j.lingua.2005.05.005.

Xu, B., J. Tanaka, and K. Mineault (2012). "The Head Turn Cueing Effect Is Sustained at Longer SOA's in the Presence of an Object Distractor". In: *Vision Sciences Society Annual Meeting*.

Yue, T., Y. Jiang, C. Yue, and X. Huang (2017). "Differential Effects of Oxytocin on Visual Perspective Taking for Men and Women". In: *Frontiers in Behavioral Neuroscience* 11. DOI: 10.3389/fnbeh.2017.00228.

Yurovsky, D. and M. C. Frank (2017). "Beyond Naïve Cue Combination: Salience and Social Cues in Early Word Learning". In: *Developmental Science* 20.2, e12349. DOI: 10.1111/desc.12349.

Ziatas, K., K. Durkin, and C. Pratt (1998). "Belief Term Development in Children with Autism, Asperger Syndrome, Specific Language Impairment, and Normal Development : Links to Theory of Mind Development". In: *Journal of Child Psychology and Psychiatry* 39.5, pp. 755–763.

Zuidema, W. (2003). "How the Poverty of the Stimulus Solves the Poverty of the Stimulus". In: *Advances in Neural Information Processing Systems*.

Zuidema, W. and B. de Boer (2009). "The Evolution of Combinatorial Phonology". In: *Journal of Phonetics* 37.2, pp. 125–144. DOI: 10.1016/j.wocn.2008.10.003.

Zwickel, J. (2009). "Agency Attribution and Visuospatial Perspective Taking". In: *Psychonomic Bulletin & Review* 16.6, pp. 1089–1093. DOI: 10.3758/PBR.16.6.1089.

Zwickel, J., S. J. White, D. Coniston, A. Senju, and U. Frith (2011). "Exploring the Building Blocks of Social Cognition: Spontaneous Agency Perception and Visual Perspective Taking in Autism". In: *Social Cognitive and Affective Neuroscience* 6.5, pp. 564–571. DOI: 10.1093/scan/nsq088.